Credit Scoring:

Aplicação da Regressão Logística vs Redes Neuronais Artificiais na Avaliação do Risco de Crédito no Mercado Cabo-Verdiano.

por

Danilson Pedro da Veiga Semedo

Dissertação apresentada como requisito
parcial de obtenção do grau de

Mestre em Estatística e Gestão de Informação

Pelo

Instituto Superior de Estatística e Gestão de Informação
da
Universidade Nova de Lisboa

Credit scoring:

Aplicação da Regressão Logística vs Redes Neuronais Artificiais na Avaliação do Risco de Crédito no Mercado Cabo-Verdiano.

Orientador: Professor Doutor Fernando José Ferreira Lucas Bação

Novembro de 2009

À minha família

e à Telma

Agradecimento

Gostaria de expressar a minha gratidão ao meu orientador Professor Doutor Fernando José Ferreira Lucas Bação, pelo apoio e amizade.

Ao Dr. Luís António Ribeiro Chorão, pelos ensinamentos econométricos que me apresentou desde os tempos de licenciatura, o que fez com que criasse curiosidade intelectual e vontade própria de me dedicar à temática do *credit scoring*, bem como pela disponibilidade e sugestões com as quais enriqueceu a presente dissertação.

Fico igualmente reconhecido ao Dr. Emanuel de Jesus Miranda, e ao Banco Comercial do Atlântico pela possibilidade que me conferiram em aceder à base de dados de crédito ao consumo, bem como à Lucília Semedo, pela disponibilidade e prontidão no esclarecimento de dúvidas relativas à base de dados.

Um agradecimento especial à Fundação Cidade de Lisboa e ao Banco de Portugal patrocinadores deste mestrado.

A todos os amigos do ISEGI, em especial ao Ricardo Vinhas, Nuno Trezentos, Elisabete Paulo, Célia Correia e João Paulo Segundo.

São, também, extensíveis os meus agradecimentos ao Dr. João Remígio e ao Dr. André Melo, pelos momentos de discussão que me proporcionaram em torno do *credit scoring.*

À Telma, pela companhia e paciência que demonstrou ao longo dos muitos dias de estudo e trabalho.

Aos meus Pais pelo apoio, coragem, incentivo e paciência que teve ao longo destes últimos tempos, a fim que este projecto se tornasse realidade.

"... O negócio de um banco é o risco! Há que reconhecê-lo, mensurá-lo e, não sendo possível bani-lo, é mister controlá-lo!".

(Chorão 2005, 121)

Resumo

A gestão de risco de crédito é sem dúvida uma das áreas mais importantes no domínio da gestão de risco financeiro. Com a recente crise financeira, e as alterações a nível da regulação introduzidas pelo acordo de Basileia II, a análise do risco de crédito e a gestão do risco em geral, têm recebido ainda mais atenção pela indústria financeira. A capacidade de discriminar bons e maus clientes tornou--se um factor decisivo para o sucesso das empresas que operam na indústria do crédito, impulsionando-as a agir de acordo com um processo de avaliação de risco mais fino. Nos países desenvolvidos, os modelos de *credit scoring* têm sido largamente utilizados neste sentido. Contudo, em Cabo Verde estas técnicas ainda estão numa fase embrionária. As instituições financeiras continuam a utilizar mecanismos indirectos de gestão de risco. Neste ambiente, alguns bancos têm procurado alinhar se às melhores práticas internacionais de gestão de risco. Neste trabalho são apresentadas duas ferramentas para a elaboração de modelos de *credit scoring* aplicado a uma base de dados de crédito ao consumo de uma grande instituição financeira Cabo-Verdiana. Depois da fase de preparação do dos dados e definida todos os parâmetros (definição da target, amostra de desenvolvimento e período de classificação), foram estimados vários modelos *logit* e várias redes neuronais multicamadas segundo diversos subconjuntos de treino/teste formados. Por fim o desempenho dos modelos é comparada com base em medidas comummente utilizados na avaliação de modelos de risco de crédito para eleger o modelo que melhor se ajusta à realidade Cabo-Verdiana. Apesar dos testes estatisticos indicarem que os modelos apresentam desempenhos estatisticamente semelhantes, as redes neuronais representam uma promissora técnica para a análise e concessão de crédito em Cabo Verde.

Palavras-chaves: *Credit scoring,* modelo *logit*, curva ROC, redes neuronais multicamadas.

Abstract

Credit risk managment, is undoubtely one of the most important area in the field of financial risk management. With the recent regulatory changes introduced by the Basel II, the credit risk analysis in particular and risk management in general, have received more attention by the financial industry. The capacity to discriminate between good and bad accounts has become a key decision factor for the success of the credit industry companies, empeling them to act according to a more sophisticated risk management process.

In developed countries, credit scoring has beem widely used in this sence. However, in Cape Verde, these techniques are still in its infancy. Financial institutions continue to use indirect mechanisms of adjudication process based on credit analyst feelings. In this environment, some banks have sought to align itself with international best practice risk management by introducing more accurate evaluation of credit. This work consists in comparing two diferent tools for the elaboration of a credit scoring model applied to a credit consumer database from a big financial instituition of Cape Verde. After database preparation and definition of the project parameter (default, sample window, performance windows) various logit models and several multilayer perceptron networks were estimated, according to different train/test subgroups formed. Finally, the performance of the models are compared based on measures commonly used to evaluate models of credit risk to elect the model that best fits the reality of Cape Verdean. Despite the statistical tests indicate that the models show statistically similar performances, neuronal networks represent a promising technique for credit adjudication process in Cape Verde.

Palavras-chaves: *Credit scoring,* modelo logit, curva ROC, redes neuronais multicamadas.

Índice

Capítulo I

1 Introdução

A gestão de risco representa um dos principais problemas enfrentado pelas instituições financeiras, desde o inicio da sua actividade. Isso ocorre, porque os bancos e as instituições financeiras em geral, têm como principal função a intermediação financeira.

No desenvolvimento da sua actividade de intermediação financeira, estão sujeitas a uma série de riscos, designadamente quando realizam operações que envolvem activos, passivos e elementos extrapatrimonias. Em relação às operações de crédito, o banco concede crédito a outros agentes económicos, sob a promessa de um recebimento futuro do capital mutuado e juros de acordo com o plano de reembolso contratado. Existe, contudo na carteira de crédito da instituição, mutuários que podem não vir a cumprir as obrigações monetárias contratados implicando prejuízos que terão de ser cobertos com as necessárias provisões. A este não cumprimento das responsabilidades por parte do solicitante de crédito chama-se de *default*[1].

Nos últimos anos devido sobretudo a pressões regulamentares, as instituições financeiras têm procurado criar metodologias mais eficientes para aferir a probabilidade de incumprimento esperado em cada operação de crédito. Contudo, só recentemente, com a crise do crédito *sub-prime* hipotecário nos Estados Unidos e, a consequente crise do mercado de crédito mundial, os consumidores, instituições financeiras e supervisores se aperceberam efectivamente da sua importância.

No âmbito internacional, tem ocorrido de forma cada vez mais acelerada, uma revolução na forma como as instituições financeiras têm avaliado o incumprimento, através de desenvolvimento de modelos internos de quantificação de risco.

[1] De acordo com o novo acordo de Basileia II, considera-se que um indivíduo está em situação de *default* desde que apresente um atraso superior a noventa dias no pagamento das prestações.

Em Cabo Verde, o processo decisório é essencialmente intuitivo, estruturando-se no "*feeling*" e na experiência dos analistas de crédito. Habitualmente são analisadas variáveis, tais como a taxa de endividamento (rácio entre os custos mensais do agregado familiar e o respectivo ordenado líquido; bens móveis e imóveis do agregado; Profissão; Tipo de contrato de trabalho; estabilidade no emprego averiguável pela antiguidade na entidade patronal; nível dos saldos médios nas contas bancárias do cliente; Entrada inicial face ao valor de preço de venda ao público e idades dos proponentes. Tendo em conta estes parâmetros, os analistas, recorriam-se ao seu "*savoir faire"* para ponderar os prós e os contras, colocando-os numa "balança mental" para avaliar o risco de crédito, isto é, para calcular a probabilidade *de default* do cliente. Recentemente, começa-se assistir a introdução de modelos de *scoring* genéricos para aquilatar se um determinado indivíduo tem perfil de bom ou mau pagador.

O aumento da concorrência entre as instituições financeiras e a crescente pressão para a maximização das receitas impulsionam as instituições financeiras, a procurarem mecanismos mais eficientes de atrair novos clientes com baixo perfil de risco e ao mesmo tempo controlar e minimizar as perdas. O aparecimento de novas tecnologias, o aumento da procura por crédito, bem como por uma questão de qualidade de serviço a necessidade de responder o mais rápido possível às solicitações levou ao desenvolvimento e aplicação de sofisticados modelos estatísticos na gestão de risco de crédito, designados por *credit scoring*.

Os modelos de *credit scoring* são sistemas que atribuem *scores* às variáveis de decisão de crédito de um requerente, mediante a aplicação de técnicas estatísticas. Esses modelos visam sumariar todas as características que permitem distinguir os bons dos maus empréstimos (Lewis, 1992). A partir de uma equação estimada com base nas características dos solicitantes de crédito, é gerado um *score* que representa o risco de perda de cada operação. O *score* que resulta da equação, é interpretado como probabilidade de incumprimento que comparado com um *cut-off* préviamente estabelecido associado a um conjunto de regras e filtros, permite ajuizar quanto à concessão ou não de crédito. Assim, a idéia básica dos modelos de *credit scoring* é identificar certos factores chave que influenciam a probabilidade de incumprimento dos clientes, permitindo a

classificação dos mesmos em grupos distintos e como consequência, a decisão sobre a aceitação ou não da proposta em análise.

Os métodos usados em *credit scoring* incluem várias técnicas estatísticas e de investigação operacional, sendo as mais utilizadas a regressão logística, a análise discriminante e as árvores de decisão (Chorão 2005). Recentemente perante o advento das novas tecnologias (aumento da capacidade de processamento) e, ao aparecimento de softwares estatísticos nos anos 80, assistimos a adopção de técnicas de inteligência artificial, como as redes neuronais e os *expert systems* (L. C. Thomas 2009).

1.1 Motivação e relevância do trabalho

A concessão de crédito desempenha um papel fundamental no desenvolvimento de uma economia, em decorrência da dinâmica que introduz no processo económico, seja como uma oportunidade para as empresas (especialmente as pequenas e médias empresas) aumentarem os seus níveis de produção ou como estímulo ao consumo dos indivíduos.

Segundo (Baptista 2006), o reconhecimento de que os mercados financeiros, através do negócio de crédito privado, contribuem para o desenvolvimento económico, é bem marcante na literatura financeira, desde (Schumpeter 1911) até (Levine 1997). A título de exemplo, o mercado de crédito ao consumo nos Estados Unidos tem demonstrado que estabilidade económica baseada em políticas sólidas de crédito é sinónimo de prosperidade económica, baixas taxas de desemprego e baixas taxas de juro. Ao longo das últimas décadas o crédito ao consumo nos Estados Unidos tem crescido num ritmo fenomenal tendo atingido em 2007 a marca de $13 triliões, superando em 40% o crédito concedido ao sector industrial e, em 24% ao crédito às empresas (L. C. Thomas 2009). A par de outros factores, o *credit scoring*, dado o automatismo que assegura foi o factor que mais permitiu a abertura do mercado de crédito a todos os consumidores, mantendo o risco num nível controlável.

Em Cabo Verde a indústria do crédito é bem menor à dos países desenvolvidos, todavia, o crédito ao consumo vem apresentando altas taxas de crescimento ao longo dos

últimos anos. Segundo dados do Banco de Cabo Verde[2] o crédito ao sector privado representa 45% do total do crédito concedido tendo registado em 2007 um crescimento de 30% face a 2006. Outros indicadores tais como o aumento expressivo da aceitação e utilização dos cartões de crédito e, o volume de transacções, associados ao facto de ter uma população maioritariamente jovem, bem como a alteração dos padrões de vida e o aumento verificado na procura por crédito a habitação, oferece um enorme potencial de crescimento do mercado de crédito ao consumo no país, quando comparado com as tendências globais. Contudo, se não existirem metodologias eficazes de previsão de incumprimento esperado e, controlo do processo de concessão de limites, as mesmas operações de crédito podem levar a economia a um processo de abrandamento, em decorrência de retracções das fontes financiadoras. Assim, para fazer face ao esperado desenvolvimento que Cabo Verde ainda conhecerá e assegurar um crescimento sustentado do mercado de crédito ao consumo, é imprescindível sistemas de *credit scoring* que permitam aos bancos e instituições financeiras avaliar automáticamente os riscos assumidos na concessão do crédito.

Ademais, este trabalho justifica-se pela crescente importância e actualidade dos modelos de *credit scoring* resultante das alterações a nível da regulação introduzidas pelo acordo de Basileia II[3].

Com este estudo pretende-se contribuir para o processo de gestão de risco de crédito em Cabo Verde, caracterizado por carentes instrumentos de avaliação e controlo do risco de crédito.

[2] Boletim Económico Banco de Cabo Verde Fevereiro 2009.

[3] Bassileia II assenta em três pilares:

Pliar I: Calculo do capital regulamentar de acordo com o *rating* das contrapartes ou de estimativas internas de probabilidades de default (PD), severidade da perda (*loss given default*, LGD) e o valor da exposição em caso de incumprimento (*Exposure at default*, EAD).

Pliar II: Análise da adequação do capital resultante da aplicação das fórmulas pré-definidas com a intervenção da autoridade de supervisão.

Pilar III: " Disclosure" da informação de gestão baseado no risco.

1.2 Objectivos

O fenómeno de *credit scoring* é ainda pouco conhecido, no caso específico de Cabo Verde. Perdura ainda uma lacuna em termos de investigação científica sobre a matéria, uma vez que, grande parte das instituições que operam no mercado não dispõe de informação sistematizada e com antiguidade suficiente que sirva de suporte ao desenvolvimento de modelos de *scoring*.

Com efeito, muitas questões se nos levantam:

- A informação de incumprimento existente em Cabo Verde é suficiente para desenvolver um modelo de *credit scoring* robusto?
- Que técnicas de desenvolvimento de modelos de *credit scoring* melhor se ajusta à realidade de Cabo Verde?

A presente dissertação tem por finalidade elaborar um modelo de *credit scoring* baseado num modelo econométrico e um modelo gerado a partir das redes neuronais artificiais para avaliação de risco de crédito relativo a solicitações de crédito ao consumo.

1.3 Organização da dissertação

Esta dissertação desenvolve-se ao longo de seis capítulos. O conjunto de objectivos propostos anteriormente traduz, ainda que parcialmente, o modo como o trabalho foi estruturado. Nesta secção ao apresentar a organização da dissertação, pretende-se orientar o leitor nas linhas seguidas ao longo do seu desenvolvimento.

Assim, após uma introdução, o segundo capítulo, apresenta uma perspectiva histórica dos modelos de *credit scoring,* a sua filosofia de funcionamento, a sua aplicação em diferentes fases do ciclo de vida de uma operação de crédito e, a sua relação com os objectivos de negócio. Apresenta-se ainda, os métodos utilizados na sua elaboração bem como as suas vantagens e limitações. Por fim, faz-se uma breve revisão da actividade de crédito em

Cabo Verde, os condicionalismos ao seu desenvolvimento e os benefícios da introdução do *credit scoring* no mercado de crédito em Cabo Verde.

O capítulo 3 dedica-se à temática da qualidade da base de dados. Começa por descrever a base de dados considerada na elaboração da dissertação, desde a selecção da janela de amostragem e o respectivo período de classificação, passando pelo processo de preparação dos dados, indivíduos considerados na modelação e selecção das variáveis. Os capítulos 4 e 5 apresentam as duas metodologias consideradas na dissertação.
Por fim, são apresentadas algumas conclusões gerais sobre o trabalho realizado.

Capítulo II

2 Credit Scoring

Desde 1960 *credit scoring* tem revolucionado profundamente os processos de decisão de crédito. O seu sucesso deveu-se em grande parte ao advento dos computadores que alterou completamente o "*Back-office"* das Instituições financeiras (Raymond, 2007).

2.1 História do *credit scoring*

Em 1936, o Estaticista Inglês, Ronald Aymer Fisher publicou um artigo sobre a utilização da técnica denominada de "Análise discriminante linear" para classificar diferentes espécies de flores do género Íris: Íris setosa, Íris versicolor e Íris verginica, analisando o comprimento e largura das sépalas e pétalas. O trabalho de Fisher forneceu as bases de análise estatística multivariada que veria a ser utilizado posteriormente em vários problemas de classificação mormente *credit scoring.*

Em 1941, David Durand no seu estudo para *National Bureau of Economic Research* (EUA), demonstrou que a mesma técnica poderia ser utilizada para discriminar bons e maus empréstimos. Segundo (Johnson, 2004) o estudo analisa 7200 observações de bons e maus empréstimos relativos a 37 empresas baseado na informação da idade, género, antiguidade no emprego, antiguidade na habitação, profissão, sector de actividade, contas bancárias, seguros de vida e valor da prestação mensal.

Mais tarde, porém no mesmo ano, os Estados Unidos vê-se envolvida na Segunda Guerra Mundial e muitas instituições de crédito e de *direct mailing* começaram a enfrentar grandes dificuldades de gestão de crédito. Muitos analistas de crédito foram recrutados para serviço militar, o que provocou uma escassez de recursos humanos com "*Know-how*" adequado para a função, numa altura em que a decisão quanto à concessão de crédito era subjectiva, dependendo, sobretudo, da experiência do analista, sem haver, portanto qualquer aplicação da técnica estatística. (Lewis, 1992) refere que Henry Wells, executivo da *Spiebel Inc corporation*, foi o primeiro a recorrer às técnicas de estatística multivariada para desenvolver modelos de *credit scoring*... Alguns anos depois, por volta

do ano de 1946, o senhor Wonderlic, então presidente da empresa "*Household Finance Corporation"*, desenvolveu um "Guia de *credit scoring*". E fê-lo recorrendo igualmente às técnicas de estatística multivariada.

Apesar dos significativos progressos registados nas metodologias dos sistemas de *credit scoring*, durante a segunda Guerra Mundial e, de estar provada a sua importância, dois factores inibiram desde logo a sua adopção: primeiramente, a resistência organizacional em utilizar os computadores no processo de decisão, e em segundo lugar, a complexidade dos algoritmos e a dificuldade de implementação dos modelos nos postos de trabalho... Mas era só uma questão de tempo!

Em 1956, com a fundação da primeira consultora na área, pelo matemático, Bill Fair e pelo engenheiro Earl Isaac, o *credit scoring* torna-se efectivamente um factor significativo na indústria do crédito. Inicialmente criaram um sistema de "*biling"* para a gestão de cartões de crédito do grupo Hilton Hotels. Dois anos mais tarde introduziram o conceito de *credit scoring,* e em 1958, produziram o primeiro modelo de *scoring* aplicacional. Uma vez que permitiam a avaliação em massa, são as empresas ligadas ao *direct mailing* e grandes cadeias de distribuição seguidas das de *leasing* que, primeiramente, utilizaram o conceito de *credit scoring* (Chorão 2005).

Durante a segunda metade dos anos 60, as empresas petrolíferas incorreram em enormes perdas devido a problemas com a gestão das operações de crédito, nomeadamente, o aumento estrutural dos eventos de incumprimento e roubos de cartões de crédito. Em resposta implementaram modelos de *credit scoring*. Nesta altura, os cartões eram emitidos sem anuidades, o que provocou, por um lado, um aumento significativo de pessoas a recorrem ao crédito, e por outro, aumento da concorrência. Muitos dos emissores de cartões de crédito de então, eram confrontados com grandes volumes de solicitações e experimentaram avultosas perdas. (Lewis, 1992) conclui que, este facto constitui a principal razão associada à introdução dos modelos de *scoring*, pelo controlo que assegura sobre a carteira de crédito.

O sucesso de *credit scoring* não foi imediato. O facto dos modelos estatísticos removerem toda a intervenção humana no processo de decisão não inspirava confiança em muitos adeptos da avaliação manual (tradicional). Apesar das resistências, *credit scoring* foi ganhando aceitação e, afirmou-se definitivamente em 1974 aquando da implementação do *Fair Credit Reporting ACT e Equal Credit Opportunity Act* em

1975/1976. Este facto tornou a discriminação ilegal no processo de atribuição de crédito, a não ser que empiricamente provado e validado estatisticamente.

Na altura, a única restrição, era a capacidade de processamento necessário. Os computadores de então, *IBM 7090 mainframe*, eram grandes, pouco eficientes comparado com os padrões actuais. Pois que, só conseguiam processar simultaneamente 26 variáveis num conjunto de 600 observações (Meys e Forgey 1963).

O sucesso do *credit scoring* na concessão de cartões de crédito nos anos 80, fez com que as instituições bancárias o aplicassem a outros tipos de bens, como o crédito pessoal, automóvel e hipotecário e, desta forma verem aumentados os seus lucros.

2.2 Filosofia de *credit scoring*

A previsão do risco e, o *credit scoring* em particular é uma área que mais desenvolvimento tem conhecido em finanças nos anos mais recentes. A par da gestão do *portfolio*, *princing options,* (e outros instrumentos financeiros), *credit scoring* representa uma importante ferramenta de estimação e redução do risco de crédito.

Na extensa literatura existem várias definições de *credit scoring*. Por exemplo (Lewis 1992) define *credit scoring*, como "um processo em que a informação sobre o solicitante é convertida em números que de forma combinada forma um *score.* Este *score* representa o perfil de risco do solicitante"; (Mester 1997) acrescenta que é um método estatístico usado para prever a probabilidade de um solicitante entrar em incumprimento. Usando dados históricos, o *credit scoring* isola as características dos solicitantes que entraram em situação de *default* produzindo, então, um *score* que a instituição utiliza para classificar o candidato ao crédito em termos de risco" (p. 3) e, decidir quanto à concessão do crédito.

Utilizado inicialmente como uma poderosa ferramenta de suporte à decisão crédito, (crédito à habitação, automóvel, cartões de crédito, crédito clássico, e crédito a pequenas e médias empresas) o *credit scoring,* é actualmente, usado para gerir e monitorizar o risco de incumprimento de todo o *portfólio* de crédito de uma instituição financeira, incluindo empresas, autoridades locais, e empréstimos especializados (*Project*

finance e imobiliária comercial). Hoje em dia não são usados exclusivamente no processo de decisão de crédito, têm tido aplicabilidade em diversos contextos como, o *pricing,* provisões, capital económico/regulamentar e titularização, como mais adiante se explicará.

Dado o sucesso dos modelos de *scoring* aplicacionais na indústria do crédito dos nossos dias, as instituições financeiras começaram a aplicá-los a outras áreas do negócio. Os modelos de *scoring* aplicacionais/reactivos têm como objectivo, determinar o perfil de risco de um novo solicitante de crédito no momento da análise da proposta. Porém, a gestão do risco de uma operação de crédito, não se resume à avaliação do risco inicial (risco no momento da análise). Importa igualmente, monitorizar o risco de crédito em toda a sua amplitude. Este acompanhamento é normalmente feito recorrendo a outro conjunto de modelos de *scoring*. Por exemplo, temos modelos conhecidos na literatura como modelos comportamentais, onde se estima a propensão à aquisição de um determinado bem (nomeadamente os modelos de *response scoring*[4] associados normalmente ao marketing); modelo de retenção, (*attrition/churn*), onde se procura identificar os clientes com maior probabilidade de abandonar a instituição.

À medida que o mercado de crédito se desenvolve, verifica-se que os modelos de *scoring* têm sido caracterizados por uma crescente sofisticação de algoritmos. (L. C. Thomas 2009) refere que estamos na era da terceira geração dos modelos de *credit scoring*, denominados por *profit scoring,* onde se pretende avaliar não só o perfil de risco do solicitante de crédito, mas igualmente, a probabilidade do candidato ao crédito dar lucro à instituição, não sendo a avaliação apenas baseada no risco. O resumo dos diferentes modelos de *scoring* usados nas diversas fases do ciclo de vida de uma operação de crédito é apresentado na figura 2.1.

[4] A aquisição de novos cliente é um processo muito dispendioso, especialmente as campanhas de *mailing.* Neste caso as instituições financeiras recorrem frequentemente aos modelos de *response scoring* para restringir o *mailing* aos clientes com maior propensão (probabilidade) de vir a responder a uma determinada campanha e portanto resultar numa relação lucrativa para a empresa.

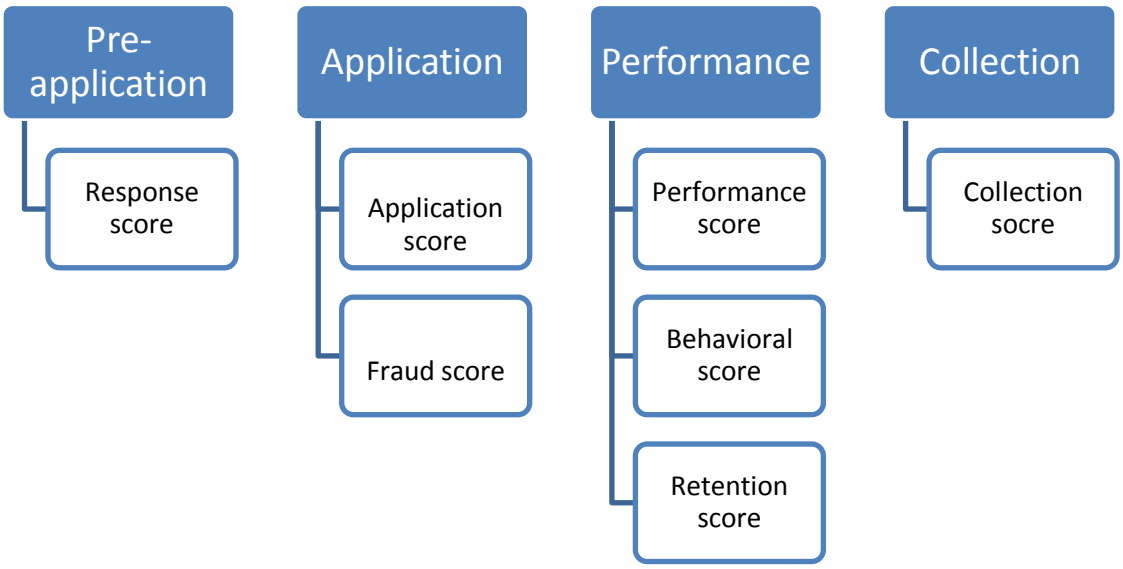


Figura. 2.1 – Modelos de scoring usados em diferentes fases do ciclo de vida de uma operação de crédito.

Fonte: Adaptado de (Gestel e Baesens 2009)

2.2.1 Scoring versus objectivos de negócio

As técnicas de *scoring* são aplicadas tendo em vista diferentes objectivos de negócio. O objectivo principal do *scoring* é melhorar o processo de selecção de bons clientes de modo a reduzir perdas futuras. Dado o seu sucesso, os sistemas de *scoring* tornaram-se um factor de decisão chave, ou se quisermos uma ferramenta de suporte à decisão imprescindível na quantificação e gestão do risco.

Os *scores* gerados pelo modelo são utilizados para calcular a perda máxima (*expected loss*) da carteira de crédito de uma instituição e, consequentemente, determinar o nível de provisões necessário para cobrir a perda máxima esperada. Para determinar a perda esperada, o risco de incumprimento da carteira de crédito precisa ser adequadamente quantificado e os *scores* têm demonstrado ser um importante *input.* Os *scores* são igualmente utilizados para determinar o montante de capital necessário para proteger as instituições financeiras e os depositantes de perdas inesperadas - capital económico/regulamentar.

Outra área recente de aplicação do *credit score* é *o pricing: risk-based pricing* (também denominado de *risk-adjusted pricing*) consiste em determinar o preço do produto tendo em conta o perfil de risco do cliente, dado pelo *score* que lhe está associado.

Muitas instituições financeiras utilizam-no igualmente para segmentar a carteira de crédito em classes homogéneas de risco e vender a investidores terceiros, como forma de redução do risco. Este processo designa-se por titularização.

Algumas instituições não financeiras têm utilizado o *credit score*, e mais especificamente os *bureau score,* para melhorar os seus processos de decisão. São exemplos, as empresas de telecomunicações e as de electricidade.

Na presente dissertação pretende-se abordar os modelos de *credit scoring* sob o ponto vista aplicacional. Assim o uso da terminologia *credit scoring* deve ser entendido neste âmbito.

2.3 Métodos utilizados em *credit scoring*

"The tools of credit scoring are based on statistical and operational research techniques and are some of the most successful and profitable aplications of statistics theory in the last 20 years".

Crook, Edelman, and Thomas (1992)

Nesta secção apresenta-se um resumo dos principais métodos paramétricos e não paramétricos utilizados em *credit scoring*, focando-se particularmente na regressão logística e nas redes neuronais. A tabela que se segue apresenta um resumo dos métodos utilizados:

Modelo	Principais Técnicas	Resumo
Regressão Linear	Minimos quadrados ordinários	Adequada em situações em que a variável resposta é contínua
Análise discriminante	Distância de Mahalanobis	Classifica os objectos em grupos pré-definidas,minimizando a variância
Regressão Logística	Estimadores de máxima verossimilhança	Adequada em situações em que a variável resposta é binária
Árvores de decisão	Chaid	Utiliza a estrutura da árvore para maximizar a variância entre-grupos
Redes Neuronais	Perceptrão multicamada	Técnica de inteligência artificial. Os resultados são difícies de explicar.
Programação Linear	Método simplex	Muito utilizado na optimização de alocação de recursos

Tabela 1- Resumo das técnicas estatísticas usadas em *credit scoring.*

Fonte: Adaptado a (Raymond 2007) (p 163)

Ao longo dos anos têm sido propostas muitas abordagens do domínio do *credit scoring.* Cada uma com as suas virtudes e defeitos, dependendo em primeiro lugar da informação disponível (base dados utilizada) e, em segundo lugar dos aspectos relacionados com a modelação.

A regressão logística, a programação linear e a análise discriminante são os métodos mais utilizados. Ainda que a maioria dos métodos apresente níveis de desempenho semelhantes[5], tem sido feito um esforço por parte dos investigadores no sentido de encontrar o método que melhor serve os propósitos de *credit scoring.* No entanto, qualquer que seja a técnica utilizada é correcto afirmar que a indústria financeira pretende em primeiro lugar modelos que tenham interpretabilidade e transparência e, em segundo lugar modelos que sejam facilmente implementáveis (Chorão 2005). A facilidade de

[5] (Thomas e N. Crook 2002), Apresenta um bom resumo de vários estudos comparativos, mostrando que existem mais semelhanças que diferenças.

implementação foi determinante na escolha do método a utilizar nos primeiros modelos de *credit scoring* desenvolvidos nos anos 1950 e 1960. Daí que a programação linear e a análise discriminante foram os primeiros métodos utilizados ainda que estatísticamente imperfeitos.

Com avanço no domínio da informática (aumento da capacidade de processamento dos computadores) foi possível testar novas abordagens como os estimadores de máxima verossimilhança. Primeiramente com os modelos *logit* (logística) e mais tarde *probit* (Gaussiana). Ambos são menos exigentes em termos de pressupostos estatísticos, mas muito exigentes computacionalmente e, inexequível numa altura em que os computadores eram tinham pouca capacidade de processamento. Hoje a diferença do tempo de processamento é incomparavelmente superior e a regressão logística é usada em mais de 80% dos modelos desenvolvidos. Devido à sua flexibilidade e facilidade de utilização os modelos de probabilidade linear[6] continuam a ser muito utilizados (Raymond 2007). Os modelos de probabilidade linear são muito utilizados em instituições onde o *credit scoring* tem uma longa história, ou onde a metodologia existente está bem enraizada. Pelo contrário, a regressão logística domina nas instituições onde o *credit scoring* foi introduzido mais tarde, quer devido às propriedades estatísticas conhecidas quer pela maior transparência e interpretabilidade que introduz no processo de decisão. Por outro lado, hoje, muitos reguladores exigem que as instituições identifiquem fortes razões para a rejeição da proposta em análise. Os modelos de *scoring* baseado na regressão logística permitem facilmente identificar estatisticamente as variáveis que mais contribuem para a rejeição do cliente.

Técnicas não paramétricas· têm sido igualmente utilizadas em *credit scoring,* com algum sucesso. Destacam-se as árvores de decisão, e métodos de inteligência artificial, como as redes neuronais, algoritmos genéticos, e método do vizinho mais próximo.

[6] A experiam por exemplo utiliza modelos de probabilidade linear nos seus modelos de *credit scoring*

2.4 Vantagens e desvantagens do *credit scoring*,

A presente secção atenta em inventariar por consulta a (Raymond 2007) as principais vantagens e desvantagens da adopção de sistemas de *credit scoring*.

A primeira vantagem da introdução de modelos de *scoring é* a redução do tempo de análise de novas propostas de crédito. Uma vez automatizado o processo, os *scores* são facilmente calculados e a resposta quanto à concessão/rejeição são geradas em tempo real, o que é extremamente importante no actual mercado de crédito cada vez mais competitivo. Este facto pode ser exemplificado pela crescente importância dos novos canais de aquisição, como a Internet, o telefone e o *E-commerce*, que faz com que o processamento e a avaliação do crédito em tempo real sejam uma necessidade.

Outra vantagem dos modelos de credit *scoring,* tem que ver com a consistência das decisões: O *score* torna o processo de decisão objectivo eliminado a possibilidade de discriminação.

É ainda possível inumerar as seguintes vantagens:

- Aumento do lucro da instituição através de maiores índices de aprovação com reduzidos níveis de *default;*
- Possibilita que o cliente seja tratado de forma personalizada independentemente do canal de entrada;
- As estratégias de risco/crédito podem ser rapidamente actualizadas e assimiladas pela organização;
- Aumenta a qualidade do serviço prestado ao cliente;
- O processo é facilmente compreendido pelos seus participantes;
- Identifica as variáveis tidas como as mais importantes na discriminação dos regulares e em situação de *default;*

A lista dos benefícios é extensa, contudo é possível enumerar as seguintes limitações.

- Custo de desenvolvimento: desenvolver um sistema de *credit scoring* acarreta custos, não somente com a instalação da infra-estrutura necessária, mas também com o suporte para a sua construção. Por exemplo: profissionais capacitados e equipamentos (*hardware / software*).
- Escassez e qualidade dos dados: Normalmente estes modelos são desenvolvidos com base nas observações presentes nas bases de dados das instituições, e que a qualidade nem sempre é salvaguardada;
- Altera a cultura da organização: a implementação desses modelos implica grandes mudanças nos processos operacionais;
- Esses modelos baseiam-se no princípio que o passado prevê o futuro, o que pode não se verificar.
- É um sistema complexo, e eventuais erros no desenvolvimento do modelo de *scoring,* definição da estratégia ou implementação podem acarretar custos para a instituição ou resultar em situações danosas na concessão do crédito;
- As técnicas estatísticas utilizadas no desenvolvimento dos modelos de *scoring* assumem que a base de dados utilizada contém um número suficiente de clientes em situação irregular. Este pressuposto nem sempre se verifica, especialmente para determinados tipos de portfólio (de pequena dimensão), onde a disponibilidade de informação é muito limitada. Para estes tipos de portfólio, é aconselhável adoptar técnicas alternativas de mensuração do risco. Ex. *experts systems,* baseados na interpretação dos C's do crédito, (capacidade, carácter, colateral, capital e condições)

2.5 Actividade de crédito em Cabo Verde

A concessão de crédito em Cabo Verde é exclusivamente efectuada pela Banca de Retalho. Considerando o universo das instituições autorizadas e em pleno funcionamento, o sistema financeiro comportava, em 2007, do lado da banca, quatro instituições de crédito, seis instituições para-bancárias e onze instituições financeiras internacionais, nove das quais dedicando-se à actividade bancária e duas a actividades de gestão de fundos mobiliários. A evolução recente na estrutura do sector bancário cabo-verdiano fica a dever-se à instalação de novas instituições, em particular, sucursais de instituições financeiras internacionais.

A dinâmica do crédito manteve-se robusta em 2008, o que pressupõe que as condições monetárias prevaleceram favoráveis à evolução da actividade económica. O Crédito Interno total apresentou uma evolução positiva em todos os meses do ano, tendo sido o ritmo de crescimento anual de 18,8%, traduzindo, sobretudo o bom desempenho do crédito à economia. O crédito bancário concedido ao sector privado ascendeu, em termos homólogos a 66.390,2 milhões de escudos Cabo-verdianos, o que representa um crescimento anual de 29,5%. A ventilação do crédito por sectores de actividade revela taxas de crescimento positivas, na sua globalidade destacando-se o crédito a particulares, que representa cerca de 53,7% do total concedido, com um crescimento de 19,2% repartidos pelo crédito à habitação (64,1%) e crédito destinados a outros fins (35,9%). O crédito líquido ao sector público administrativo, registou uma redução moderada de 8,1% quando comparada à diminuição de 25,5% ocorrida em 2007, reflectindo os efeitos da diminuição verificada nos depósitos do Governo Central junto ao Banco Central.

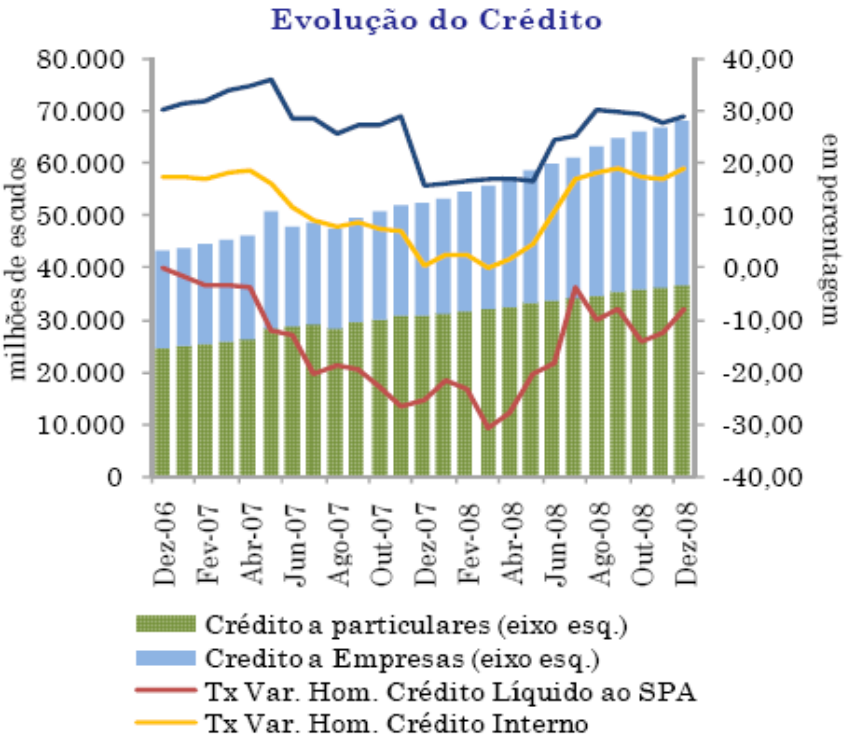


Tabela: 2.2. Evolução do crédito 2008.

Fonte: Boletim Económico Fevereiro de 2009, Banco de Cabo Verde.

Em termos da qualidade dos activos, o peso da carteira de crédito vencido dos bancos, no total do crédito, apresentou um acréscimo de 3,9% em 2006, passando para 13,5% em

2007, o que se fez acompanhar da mesma tendência pelo rácio crédito vencido líquido de provisões sobre o capital, derivado, essencialmente, da aplicação do novo regime de classificação de crédito e constituição de provisões. Contudo, torna-se premente efectuar melhorias permanentes nos sistemas de gestão e maior controlo do risco de crédito, mediante um acompanhamento contínuo da evolução do crédito mal parado e análise das suas interligações com algumas variáveis macroeconómicas relevantes.

Distribuição Sectorial do Crédito Bancário

(em percentagem do total do crédito concedido)

	2005	2006	2007
Crédito concedido a Empresas não Financeiras	**34,9**	**42,9**	**41,4**
Agricultura, Silvicultura, Caça e Pesca	1,2	0,8	0,8
Indústrias Extractivas	0,0	0,0	0,0
Indústrias Transformadoras	4,2	3,6	4,0
Electricidade, Água e Gás	0,5	11,3	4,7
Construção e Obras Públicas	3,0	4,1	5,1
Comércio, Restaurante e Hotéis	11,1	9,6	10,6
Transportes e Comunicações	5,7	5,5	4,9
Serviços Prestados às Empresas	0,8	0,9	1,9
Serviços Sociais e Pessoais	8,4	7,0	9,4
Crédito concedido a Particulares (inclui créditos a emigrantes)	**65,1**	**57,1**	**58,6**
Habitação	52,7	47,2	45,2
Outros fins	12,4	9,9	13,4
Total de Créditos Concedidos	**100,0**	**100,0**	**100,0**

Fonte: Banco de Cabo Verde
Nota: 2006 passa a incluir o crédito concedido pela CECV

Tabela: 2.3. Distribuição do crédito Bancário por sector de actividade.

2.6 Condicionantes da actividade de crédito e benefícios da introdução do *credit scoring* em Cabo verde

Após uma série de reformas, o sector financeiro Cabo-Verdiano, está cada vez mais moderno, competitivo e concorrencial, conforme indica o número crescente de agências bancárias, a melhoria da qualidade do atendimento, e a disponibilidade de novos meios de pagamento (ATMs, POS, cartões de crédito). O índice de penetração dos serviços financeiros, seja em termos geográficos, seja em termos demográficos, é indicador da evolução positiva registada pelo sistema nos últimos anos.

Porém, o desenvolvimento do sistema tem sido condicionado pelo elevado custo de intermediação financeira e altas taxas de juro, característica peculiar de sistemas financeiros de países em vias de desenvolvimento. O alto custo do dinheiro apresenta-se, assim, como um obstáculo importante para a expansão do crédito, importante factor de concentração do rendimento e da riqueza, influenciando negativamente o nível de investimento na economia.

Neste contexto, a introdução do *credit scoring,* pelas características que lhe estão associadas, introduz inúmeros benefícios no mercado de crédito Cabo-verdiano, dos quais que destacam:

- O *credit scoring* é uma parte vital do bom funcionamento de um sistema financeiro moderno permitindo a redução do foco na análise manual (tradicional) das propostas, baseados essencialmente na interpretação dos C's do crédito (carácter, capacidade, colateral, Capital, e Condições), o que traduz uma melhorar capacidade de análise dos pedidos de crédito e aferição do perfil de risco dos clientes (Turner 2006).

- Melhora o *trade-off* entre o volume de aquisições e o controlo do crédito mal parado. Como refere (Baptista 2006), o prémio de risco, representado pelo nível de provisões para perdas de crédito, constitui um factor de peso na formação do custo final de intermediação em Cabo Verde. Assim, uma melhor análise dos pedidos de crédito e acompanhamento dos clientes implica uma melhor alocação de provisões para perdas esperadas de crédito. Isso poderá implicar menor

necessidade de constituição de provisões, portanto, de recursos não produtivos que, em última medida, iriam contribuir para a redução dos custos globais da instituição e tornar mais barato o crédito aos clientes. (Mateus 2000) acrescenta, que a criação de condições para redução e melhor avaliação do risco e, para recuperação do crédito mal-parado é factor que contribui para redução do custo de financiamento. Ainda segundo o mesmo autor, o prémio de risco é um factor que acrescenta 1,85 p.p. ao *spread*.

- Melhora a eficiência operacional, quando o processo está automatizado, os *scores* são facilmente calculados e removem as tarefas demoradas da análise. Este facto pode levar a uma expansão dos níveis de crédito concedido, decorrentes do maior grau de certeza das instituições bancárias quanto às perdas nos financiamentos e maior rapidez na análise das propostas de crédito.

- À medida que a concorrência aumenta, os modelos de *credit scoring* permitem adoptar estratégias de *Risk Based Princing* (RBP), ou seja, determinar o *pricing* da operação de acordo com o perfil de risco do cliente. Este facto permite oferecer taxas de juro mais concorrenciais para clientes com baixo perfil de risco e, potencialmente maior disponibilidade de crédito para clientes de alto risco, que de outra forma viriam os seus pedidos de crédito simplesmente recusados.

Por outro lado, face às exigências da globalização e a crescente necessidade de modernização do sistema financeiro, as instituições financeiras em Cabo Verde têm pela frente um conjunto de desafios, que tornam desejável a introdução dos modelos de *credit scoring*:

- A necessidade de investir em tecnologia;
- Reduzir as taxas de juro,
- Crescente aumento da concorrência e consequente pressão para a diversificação (novos produtos).

Estes desafios, bem como a crescente procura por crédito, impulsionará as instituições bancárias a procurarem economias de escala e, a agirem de acordo com um processo de

avaliação de risco mais fino, conduzindo a uma maior consolidação da indústria de crédito em Cabo Verde.

2.7 Supervisão e gestão de risco de crédito no sector bancário em Cabo Verde.

A regulação do sistema financeiro supervisão prudencial das instituições de crédito é focalizada em aspectos relacionados com a concentração do crédito, constituição de provisões e fundos próprios, analize dos riscos para a estabilidade financeira, e avaliação da capacidade de absorção de choques no sistema.

A nível mundial, e associado ao fenómeno da globalização do mercado financeiro, tem-se verificado uma intensa movimentação de países no sentido de fazerem convergir os seus sistemas de normas e regras internas aos padrões internacionais, principalmente tratando-se de países de economias e mercados mais abertos e competitivos. Para países como Cabo Verde, de mercados pequenos e com elevado grau de dependência externa, a necessidade de estar em linha com os padrões internacionais torna-se um imperativo vital ao processo de convergência. Nesta perspectiva, e face a um sistema financeiro cada vez mais exigente, mais exposto e em fase acelerada de desenvolvimento, iniciou-se nos últimos anos em Cabo Verde um conjunto de reformas legais e de normativas prudenciais, mais consentâneas com o desenvolvimento do sector financeiro, visando proceder a uma necessária aproximação aos actuais padrões de referência internacional, designadamente, as Normas de Reporte Contabilistico e Financeiro Internacional (IAS/IFRS) e BASILEIA II.

De 1998 a meados de 2007, a gestão de riscos de crédito no sector bancário em Cabo Verde esteve ancorada ao Aviso n.º 09/98[7]. Este normativo, de natureza de gestão meramente administrativa do risco, revelou-se, após anos de vigência, incapaz de atender

[7] Aviso n.º 9/98, de 28 de Dezembro, do Banco de Cabo Verde:Estabelece o nível mínimo de provisões que as instituições sujeitas a supervisão do Banco de Cabo Verde devem observar. (BO n.º 48)

à realidade actual, que se caracteriza por: (i) aumento e complexidade de situações de risco de crédito, antes não previstos; (ii) possibilidade, capacidade e necessidade dos bancos se socorrerem de métodos e instrumentos eficazes de gestão de risco de crédito; factos que reclamam o estabelecimento de mecanismos de diferenciação entre os bancos em matéria de gestão de riscos nas suas actividades.

Com efeito, no actual estágio de evolução do sistema financeiro nacional, e não alheio à concorrência de um mercado global, o instrumento consubstanciado no Aviso n.º 09/98 demonstrava-se incapaz de corresponder às exigências e aos desafios do mercado, devido à insensibilidade e inflexibilidade que o caracterizava em relação ao risco, tendo acomodado na sua estrutura conceptual procedimentos que obrigavam a tratamentos igualitários para situações absolutamente díspares. Em resposta a essas situações, foi concebido, com a assistência técnica especializada do FMI, o novo normativo, o Aviso n.º 04/2006[8]. Este Aviso é mais sensível ao tratamento de situações de risco, ao introduzir alguns mecanismos que permitem tratar situações diferentes, pretendendo servir de ponte entre um sistema de gestão administrativa do risco de crédito e um sistema e cultura de gestão económica do risco.

[8] Aviso n.º 4/2006, de 2 de Janeiro de 2007. Estabelece a classificação de operação de crédito e provisões. Revoga o aviso n.º 9/98, de 28 de Dezembro. (B.O. nº 1, I Série).

Capítulo III

3 Caracterização da base de dados de análise

Este capítulo tem como objectivo descrever a base de dados utilizada neste estudo. Foca, ainda, nos aspectos relacionados com a preparação dos dados, identificando as principais considerações a ter em conta na construção de uma base de dados, que vão desde o tratamento dos *missing values*, passando pelas exclusões, até a definição da variável *target,* janela de amostragem e período de classificação de forma a alimentar a fase da modelação.

A base de dados utilizada neste estudo foi fornecida por uma instituição bancária Cabo-Verdiana e, como é requerido neste estudo e em casos semelhantes, foi quebrada qualquer possibilidade de identificação dos clientes nela constantes. A base de estudo é composta por 15.000 registos referentes a créditos ao consumo concedidos a clientes particulares no período de Janeiro de 2004 a Abril de 2009. Foram consideradas na análise todas as variáveis constantes no formulário de proposta de crédito em uso na instituição (tabela 3.1).

ID	Variável	Tipo de Variável
#1	Estado civil	Catégórica
#2	Género	Catégórica
#3	Profissão	Catégórica
#4	Actividade professional	Catégórica
#5	Entidade patronal	Catégórica
#6	Cargo na empresa	Catégórica
#7	Idade	Contínua
#8	Habilitações literárias	Catégórica
#9	Nacionalidade	Catégórica
#10	Naturalidade	Catégórica
#11	Antuiguidade como cliente	Contínua
#12	Rendimento mensal	Contínua
#13	Prazo do empréstimo	Contínua
#14	Taxa de juro	Contínua
#15	Valor da prestação mensal	Contínua
#16	Valor solicitado	Contínua
#17	Valor financiado	Contínua

Tabela 3.1 –Definição das variáveis

A natureza dos dados extraídos pode-se estruturar em três tipos distintos:

- Caracterização do cliente;
- Caracterização da operação de crédito;
- Comportamento do cliente.

A primeira tem que ver com a informação que caracteriza o cliente na sua esfera sócio demográfica, sendo exemplo disso a idade, profissão, estado civil, etc. A segunda tem a ver com a caracterização da operação de crédito, isto é, o bem, o valor a financiar, o prazo da operação, etc. E finalmente a terceira, prende-se com as informações respeitantes ao comportamento dos clientes, apurando mensalmente o número de prestações/dias em atraso, durante o período de vigência do contrato. Tanto a primeira como a segunda são recolhidas no momento de solicitação do crédito e, constituirão as variáveis independentes dos modelos que iremos utilizar. Por sua vez a terceira servirá, como mais a frente se explicará, para definir a variável dependente, ou seja, “Bom” ou “Mau” pagador.

3.1 Qualidade da base de dados

"Neither sophisticated software nor statistical techniques can overcome the inherent limitations of the raw data that goes into them."

(Wynn 2003)

A preparação da base de dados de análise é um estágio importante no processo de desenvolvimento de um modelo preditivo. Este estágio fica a dever-se ao facto de a maioria dos dados que podemos encontrar serem pouco adequados para os propósitos que se pretendem. Entre os problemas que normalmente encontramos contam-se exemplos de dados inválidos e inconsistentes e o aparecimento de *missing values* e *outliers*. Relativamente aos primeiros, a sua identificação e posterior remoção é importante, na medida em que a sua existência pode comprometer a validade dos resultados finais, dos quais são exemplos de cuidados a observar: atribuição de crédito a indivíduos que não tenham sido sujeitos a uma decisão baseados no sistema de avaliação de crédito ou que tenham sido excluidos por outros motivos, por exemplo, crédito a colaboradores, VIPs, indivíduos com idade inferior a 18 anos ou com histórico de elevado risco e propostas suspeitas de fraude.

Quanto aos *missing values,* é frequente encontrarmos, nas bases de dados, padrões incompletos, ou valores que não fazem sentido para uma determina variável. Existem diversas formas de lidar com este problema, não havendo uma que seja nitidamente superior a todas as outras, das quais se destacam:

1. Excluir todos os registos/variáveis que apresentam valores omissos

2. Excluir da amostra de desenvolvimento registos/variáveis que apresentam uma percentagem significativo de *missing values* (ex.50%), especialmente se for expectável que o nível de *missing values* se mantenha no futuro.

3. Considerar os *missing values* como um novo atributo das variáveis a incluir no modelo.

4. Utilizar técnicas estatísticas para preencher os campos em falta. Uma opção para a resolução do problema traduz-se no preenchimento automático dos campos com uma boa estimativa do seu valor. Existem diversas formas de produzir esta estimativa sendo que a mais simples consiste em adoptar medidas de tendência central como a média, a mediana ou a moda. Outra abordagem interessante consiste na especificação do mesmo como um problema econométrico. A ideia é desenvolver um modelo preditivo que, com base nos registos completos e nas variáveis disponíveis nos forneça uma boa estimativa para os valores em falta.

O risco que corremos ao optar pelas duas primeiras soluções é relativamente óbvio e traduz-se em primeiro lugar, na não utilização de variáveis importantes para a formulação do modelo explicativo. Poderemos estar a prescindir de variáveis que na realidade são importantes para modelar o fenómeno e sendo este o caso, o modelo provavelmente nunca produzirá resultados tão precisos quanto estariam ao seu alcance, caso as referidas variáveis de input fossem utilizadas. Segundo, o facto de determinados registos não apresentarem valores, pode em si mesmo, evidenciar um caso importante (ex. Mau *performance*). Ao excluir estes registos corremos o risco de estar a enviesar a amostra, isto caso haja um motivo para que estes registos não possuam valores para a variável em causa.

Por exemplo sabemos que a probabilidade de um indivíduo recém-empregado não preencher o campo destinado à “antiguidade no emprego” no formulário da proposta de crédito é elevado. Com efeito, caso existam uma grande percentagem de indivíduos que não preencheram o campo destinado à “antiguidade no emprego”, decidimos retirá-lo do nosso conjunto de dados. Ora, quando tal facto acontece, corremos o risco de estar a enviesar a nossa amostra. Se é certo que indivíduos recém-empregados tendem a deixar o referido campo em branco, isto quer dizer ao eliminarmos estes mesmos registos, estaremos também a retirar da amostra uma grande parte de indivíduos com poucos anos no emprego. Como consequência o nosso modelo produzirá sempre estimativas pouco precisas, senão mesmo aleatórias, para os indivíduos com poucos anos no emprego.

Pelas razões apresentadas acima e por se considerar que a inclusão dos *missing values* na amostra acrescenta benefícios ao negócio, optou-se neste estudo pela opção 3.·

3.2 Janela de amostragem e período de classificação.

A expressão "o passado prevê o futuro" representa um dos princípios fundamentais da evolução dos modelos de *credit scoring.* Baseado neste princípio, os dados históricos de anteriores solicitantes de crédito são analisados para prever o comportamento de futuros proponentes de crédito. Assim, selecciona-se um conjunto de clientes abertos num determinado período de tempo, denominado de janela de amostragem, e o seu comportamento é analisado noutro período distinto no futuro, chamado de janela de classificação, para determinar a variável dependente, isto é, se o cliente foi regular ou *default*. Mas que horizonte temporal seleccionar? Não há, na verdade, uma resposta que seja clara e objectiva quanto ao tempo a considerar na análise. A escolha da melhor amostra deve responder a dois aspectos importantes: em primeiro lugar, a informação seleccionada deve ser o mais recente possível, de modo a reflectir o perfil de futuros solicitantes de crédito. Em segundo lugar, deve cobrir um período de produção significativo, de modo a garantir um número suficiente de bons e maus contratos para a modelação. Há portanto dois objectivos conflituais: Se o período de análise é demasiado pequeno, então os indivíduos em situação de *default* serão classificados como regul ares (**erro tipo I**); se o período, pelo contrário, for demasiado longo, apesar de se terem mais observações de indivíduos em *default,* os dados estarão desactualizados e não servirão os intentos de utilização preditiva do modelo (Wynn 2003).

A literatura é escassa e não perfeitamente concludente no que toca à determinação da janela de amostragem que elege os registos que participarão na construção do modelo. Podemos contudo referir que a janela de amostragem deverá ter uma taxa de maturação estável, isto é, a taxa de *default* da carteira do produto de crédito deverá apresentar características de estabilidade ao longo do tempo. Conhecendo assim a maturidade da população, estamos em condições de seleccionar a amostra de desenvolvimento constituída pelos indivíduos cuja maturidade é igual ou superior a maturidade global e que, portanto, seguramente os poderemos classificar num de dois grupos, regular ou

default, reduzindo deste modo a possibilidade de ocorrência do **erro tipo I** como explicado anteriormente.

O momento de maturação da taxa de *default* (momento a partir da qual a taxa de *default* não evolui mais), e a amplitude dos períodos, janela de amostragem versus período de classificação, variam de produto para produto e da definição de *default* utilizada, não havendo uma regra inequívoca e explicita para a sua determinação. Segundo (Siddiqi 2006) os modelos aplicacionais apresentam características de estabilidade entre os 18 a 24 meses, enquanto se se tratar de um crédito hipotecário, a regra é de 3 a 5 anos. Contrariamente, nos modelos comportamentais, normalmente são utilizados períodos de análise mais curtos, entre 6 a 12 meses, e 1 mês ou menos para modelos de recuperação. Quando o objectivo do desenvolvimento do modelo tem a ver com questões de carácter meramente regulatório, nomeadamente no âmbito do novo acordo de Basileia II, o período de classificação é estabelecido pelo regulador (12 meses). No presente estudo assumiu-se um compromisso de 12 meses na janela de amostragem e 24 meses no período de classificação, dado garantirem observações em número suficiente para a estimação do modelo e conferirem uma estabilização da taxa de *default* como era mister encontrar, conforme anteriormente citado.

3.3 Definição de bom, mau e indeterminado.

A classificação de clientes quanto ao incumprimento é uma etapa chave do processo de desenvolvimento de um modelo de *credit scoring*. Sem dúvida que o que pode ser um "bom" cliente para uma organização, poder ser "ma" para outra dependendo da ambiente de negócio. Por esta razão (Wynn 2003, 53) menciona que a definição de incumprimento deve reflectir a experiência da própria instituição.

(Siddiqi 2006, 39) apresenta uma lista de aspectos a considerar na definição de incumprimento:

- Deve estar em linha com os objectivos da organização. Por exemplo, se o objectivo da instituição é aumentar os lucros, então, o mau cliente deve ser definido em função do conceito de rentabilidade.

- Deve estar em linha com o produto e com os intentos de utilização preditiva do modelo;
- Deve garantir por um lado, um número suficiente de observações para suportar a fase de modelação e, por outro, uma definição que seja capaz de diferenciar bons de maus clientes.
- Deve ser fácil de interpretar;
- Em algumas situações, poderá ser vantajoso ter a mesma definição de *default* (mau) em diferentes segmentos ou mesmo modelos em produção na instituição. Este facto torna mais fácil o processo de gestão de risco e consequente tomada de decisão, especialmente em ambientes onde existem vários modelos de *credit scoring*.

Porém, com a entrada em vigor do novo acordo de Basileia II, à semelhança da definição da janela de amostragem e período de classificação, a definição do *default* é também definido pelo regulador. Neste caso, considera-se que um cliente está em situação irregular (*default)*, se ultrapassar mais de noventa dias nos primeiros doze meses de vigência do contrato.

Neste estudo adoptou-se a definição de Basileia II para classificar os clientes quanto ao incumprimento.

Uma vez definido os “maus” clientes, a mesma análise efectuada anteriormente pode ser utilizada para definir o conjunto dos “bons/regulares" clientes. Novamente, esta deve estar em linha com as questões discutidas anteriormente. A definição de regular (bom) é menos analítica e muitas vezes óbvia. No presente estudo considera-se que o cliente está em situação regular se liquidou todas as prestações dentro do prazo estabelecido. Um aspecto importante a anotar é que, enquanto um cliente regular precisa manter a sua condição de “regular” ao longo da janela de classificação, um “mau” cliente só precisa atingir a definição adoptada uma única vez em qualquer altura dentro da janela de classificação.

Existe ainda um conjunto de indivíduos, que não tendo comportamento suficiente, não os poderemos classificar num de dois grupos, regular ou *default.* Não se encontram

suficientemente maduras para ter a capacidade de se ter tornado delinquentes ou mesmo para ter falhado alguma prestação. Este conjunto de indivíduos nesta situação denominam-se indeterminados e, é comum em *credit scoring,* não os considerar na modelação.

3.4 Inferência dos rejeitados

Um dos maiores problemas no processo de desenvolvimento de modelos de *credit scoring* consiste na evidência que somente as propostas de crédito aprovadas e concretizadas, poderão ser classificadas como: ”Bons”, “Maus” e “indeterminados”. Para as propostas recusadas no passado, apenas detemos as variáveis dos proponentes, mas obviamente não possuímos a informação de “Bons” ou “Maus”. Se estes clientes, recusados, forem ignorados e retirados da população de desenvolvimento, provocará um “*bias*”, quando o modelo, construído sobre os “Bons” e os “Maus”, classificar um proponente recusado anteriormente. Pois pode se deixar de avaliar algumas características específicas, que esteja particularmente presente apenas nos proponentes rejeitados, fazendo com que o novo modelo de *credit scoring* desenvolvido não consiga prever de forma adequada o comportamento desses indivíduos. De modo a incluir estes clientes no modelo, utiliza-se uma técnica denominada de inferência dos rejeitados. Esta técnica visa por um lado, inferir o comportamento dos solicitantes rejeitados no processo de decisão de crédito, e reduzir o enviesamento da selecção da amostra, por outro.

A literatura é ainda muito escassa no que a este tema diz respeito, somente encontramos uma série de estudos que avaliam de modo empírico as técnicas de inferência dos rejeitados em *credit scoring.* As técnicas de *extrapolation* e *augmentation*, aqui tratados como dados aumentados foram inicialmente propostos por (Hsai 1978), depois por (Hand e Henley 1993) e (Banasik e Crook 2005), contudo, os estudos empíricos levados a cabo por (Crook e Banasik 2004), demonstraram não haver vantagens na inclusão deste grupo no processo de estimação. (Dempster, Laird e Rubin 1977) utilizou o algoritmo de *expectation e Maximization* (EM) para a estimação da máxima verosimilhança a partir do tratamento dos rejeitados como dados incompletos; (Joanes 1993) desenvolveu um modelo de *credit scoring* com base num conjunto de solicitantes aprovados recorrendo a regressão logística, que utilizou posteriormente para inferir o comportamento dos

rejeitados. (Ash e Mester 2002) apresentaram o *parceling,* os mesmos autores sugerem ainda utilizar informação de mercado, para inferir o comportamento de solicitantes rejeitados. (Feelders 2000) considera a inferência dos rejeitados como um problema de dados omissos. E (Shin e Sohn 2006) utilizam a técnica de análise de sobrevivência, apresentando um método de inferência dos rejeitados baseados no intervalo de confiança para a mediana do tempo de sobrevivência dos clientes em *default*.

Ao desenvolver um modelo de *credit scoring*, pretende-se em primeiro lugar que este seja representativo do comportamento de todos os solicitantes de crédito. Contudo, tipicamente os modelos são desenvolvidos apenas com base em informação comportamental dos clientes aprovados, pois o comportamento dos clientes rejeitados é desconhecido. A inferência de rejeitados pode ser, então, entendido, segundo (Shin e Sohn 2006) como um processo de estimação do risco dos indivíduos rejeitados no processo de decisão de crédito.

Existem várias técnicas que utilizam os indivíduos rejeitados no desenvolvimento de modelos de *credit scoring*. Entre elas, estão as mais citadas na literatura, como: a classificação dos rejeitados como clientes maus, parcelamento (*parceling*) e dados aumentados (*augumentation*) e ainda a utilização de informação de mercado como um método de inferência dos rejeitados.

3.4.1 Parceling

Apresentado por (Ash e Mester 2002) é caracterizado como um processo de reclassificação por risco. Basicamente, o método consiste em segmentar a população dos rejeitados em clientes Bons e Maus, segundo o risco observado no conjunto de clientes aprovados. Para cada intervalo de *score* é feito uma partição aleatória dos rejeitados, com base na frequência observada de Bons e Maus, presentes na população dos aprovados. Um novo modelo será então desenvolvido a partir da nova base de dados redistribuída, ou seja, com todos os solicitantes rejeitados reclassificados como Bons e Maus clientes e adicionados à base inicial de clientes aprovados. Apenas pode ser utilizado em instituições onde existe modelos de *credit scoring* em produção, uma vez que para

efectuar a reclassificação dos solicitantes rejeitados é preciso conhecer a taxa de maus por *buckets* de *score*. Uma alternativa para a utilização deste método na ausência de um modelo de *credit scoring* consiste em efectuar a reclassificação de rejeitados de modo aleatório a partir da taxa total de maus observada na amostra de desenvolvimento (proponentes aprovados).

3.4.2 Augumentation (dados aumentados)

O método de dados aumentados é o método mais utilizado em *credit scoring* e, está disponível em muitos softwares estatísticos. É geralmente utilizado quando o processo de análise de risco de crédito é feito com base num conjunto de filtros e regras de risco. Esse método considera que para o mesmo "*score*" a probabilidade de um rejeitado/recusado ser "Bom" é igual a probabilidade, de um aprovado ser "Bom". Assim, em primeiro lugar estima-se um modelo com base nos proponentes aprovados e rejeitados (*Accepted/Rejected model*). Em seguida gera-se um novo modelo ponderado, com apenas os proponentes aprovados, (*Good/Bad model*) utilizando como variável de ponderação o peso obtido no modelo (*Accepted/Rejected model*) inicialmente desenvolvido.

3.4.3 Classificação de rejeitados como clientes maus

Uma das abordagens mais simples de tratamento dos rejeitados é classificá-los como maus clientes. Assim, a amostra de desenvolvimento do novo modelo será composta por clientes aprovados (Bons e Maus) acrescidas dos solicitantes rejeitados, todos classificados como clientes Maus. Esta técnica nada aconselhável ainda colhe adeptos actualmente.

3.4.4 Utilização de informação de mercado

Este método utiliza informações de mercado, obtidas a partir de uma central de informação de crédito para inferir o comportamento dos proponentes rejeitados, ou seja, sobre os clientes recusados numa determinada instituição financeira e se aprovados

noutra, obter informação sobre o seu comportamento no pagamento. Este método assume que o proponente comporta de maneira semelhante independentemente da instituição.

Quando utilizamos informações de mercado, temos um ganho natural de informação, para os novos modelos desenvolvidos, pois temos informações adicionais, para além das informações internas disponíveis na instituição credora. Porém a obtenção de informações de mercado junto das centrais de crédito, exige um custo, que deve ser considerado e avaliado no momento do desenvolvimento de novos modelos.

3.4.5 Potenciais benefícios da utilização da inferência dos rejeitados

Como referido anteriormente, pouco tem sido publicado sobre a temática do *reject inference*, sendo que a maioria destes focam-se em apresentar as técnicas de inferência dos rejeitados e, pouco tem sido feito no sentido de quantificar os seus benefícios. Porém, da pouca investigação disponível, parece não haver consenso. (Crook e Banasik 2004), defendem que os potenciais benefícios da introdução dos rejeitados no desenvolvimento do modelo são modestos. Por outro lado, (Siddiqi 2006) argumenta que a inclusão dos rejeitados, constitui uma mais-valia no processo de desenvolvimento de modelos de *scoring*, reduzindo o impacto do enviesamento amostral. (Montrichard 2007) demonstrou empiricamente que a inclusão dos rejeitados permite:

1. Identificar as características de clientes associados ao risco de crédito;
2. Obter estimativas mais precisas da taxa de maus;
3. Aumentar a capacidade do modelo em distinguir os bons dos maus clientes;
4. Facilita a comparação de modelos candidatos.

O segundo ponto é principalmente importante do ponto de vista de aceitação. Normalmente o *cut-off* é determinando fixando uma taxa de aceitação que confere um nível de risco aceitável, isto é, que a instituição está disposta a assumir. Ora, se a estimativa da nova taxa de maus for subestimada, a instituição incorrerá em perdas inesperadas, o que é altamente indesejável.

3.5 Selecção das variáveis

Quando se seleccionam dados no âmbito de um problema de classificação a tendência é acrescentar o maior número de variáveis possíveis, de forma a bem caracterizar o problema. Acontece, normalmente, que muitas das variáveis pouco ou nada estão associadas ao conceito-alvo, (*target*), havendo nestes casos dois tipos de variáveis: As variáveis completamente irrelevantes, ou seja, que em nada distinguem o conceito-alvo e as variáveis redundantes, ou seja, que em nada acrescentam a discriminação do conceito-alvo. Por esta razão, é comum em estudos deste genêro, levar a cabo diversas abordagens de forma a encontrar as relações tidas como as mais preditivas para o objectivo em estudo.

O propósito da selecção de atributos consiste em, a partir de um conjunto inicial de **F** atributos, seleccionar um subconjunto **G**, tal que **G** < **F**, tendo sido **G** apurado segundo um determinado critério que permita identificar as variáveis relevantes para o problema em análise. A eliminção de atributos inúteis permite reduzir a dimensão dos dados de treino e a sua complexidade e, portanto, reduzir o tempo de processamento dos métodos a aplicar nas fases seguintes. Além disso, (Hosmer e Lemeshow 2000) apela para a importância da selecção de variáveis, pois tendencialmente, com um menor número de variáveis o modelo será mais robusto. Este assunto, muito querido dos estatísticos, e válido, pois quanto maior o número de graus de liberdade subjacentes ao modelo, maior será a dependência do modelo ao conjunto de treino e, portanto maior a sua variabilidade.

A selecção de atributos deverá eleger o subconjunto de atributos, com maior relêvancia para o conceito-alvo, não perdendo de vista duas condições: A primeira é o de a capacidade preditiva do modelo não diminuir significativamente. A segunda e que as probabilidades condicionadas P (x|bom) e P (x|mau), que representam as funções de densidade de probabilidade para cada um dos grupos, “Bom” e “Mau” se mantenham semelhantes, para todos os elementos de ambos os grupos, antes e depois da selecção de atributos. Foram, assim, encetadas várias análises, todas com uma mecânica comum que se sintetizam, basicamente, a aspectos de índole gráfica e estatística:

Primeiramente, efectuamos uma análise bivariada sobre os dados da janela de amostragem a fim de aferir a capacidade discriminatória de cada variável, na construção do modelo.

Seguidamente, outro tipo de análise para exploração dos dados diz respeito ao cálculo dos *odds e* dos *odds-ratio.* O *odds* pode ser interpretado como a comparação de dois números: "o primeiro traduz a probabilidade de ocorrerência de um evento; o segundo, a probabilidade do mesmo evento não ocorrer". Ou matemáticamente;

$$Odds = \frac{P(evento)}{1 - P(evento)}$$

Por fim, atendento à comparação que se pretende efectuar na identificação de quais os atributos que deverão constar num determinado modelo de *scoring*, calcula-se o rácio entre os *odds,* isto é, o *odds-ratio* (*OR*).

$$OR = \frac{odds(Y = 1 | X = 1)}{odds(Y = 1 | X = 0)}$$

Por outras palavras, o *odds-ratio*, é uma medida de associação que indica o quanto mais ou menos provavél é a probabilidade de obter uma resposta positiva, consoante o valor da variável independente. Por exemplo para variáveis explicativas dicotómicas, se considerarmos que Y indica se o indivíduo está em situação regular ou em *default,* e X presença ou ausência de uma determinado factor de risco (característica do indivíduo), então o *odds-ratio* indica-nos o quanto mais provavél é a ocorrência do evento, neste caso, *default*, consoante o factor de risco está ou não presente. Um *Odd-ratio* igual a 1 indica ausência de relação entre a variável explicativa e a dependente; um OR menor que 1, indica que a variável explicativa está negativamente associado à *target*, ou seja, quanto menor o *odds-ratio,* maior é a probabilidade de o cliente apresentar menores risco de incumprimento, indicando que o factor de risco apresenta algum poder para disciriminar clientes bons. Um OR> 1 significa que quanto maior é *OR*, maior é a probabilidade de o cliente apresentar maiores riscos de incumprimento, evidenciando que o factor de risco considerado apresenta poder para discriminar maus clientes.

Outro estudo preliminar consiste em agrupar e discretizar os atributos a fim de poderem explicitamente, estar espelhados no modelo. (Sarmento 2005, 46) Apresenta várias razões pelas quais a dicretização se torna muitas vezes indispensável: Em primeiro lugar, se um atributo numérico, possuir valores omissos, uma das formas será discretizar o atributo, para que se possa tratar o *missing* como um novo atributo. Em segundo lugar, nos problemas em que as regressões lineares são utilizadas, a discretização é um meio importante para fornecer robustez ao modelo resultante, tornando-o mais generalizável. A discretização é também um meio para combater os valores extremos e os "*outliers*" que tanto perturba a estimação dos parâmetros. No entanto a discretização, não é gratuita, faz-se à custa de perda de informação do atributo, mas em nome da abstracção. O problema está em como discretizar optimizando o binómio perda de informação versus abstracção. A este propósito (Thomas e N. Crook 2002) considera necessário a agregação de atributos pois que "há, normalmente um grande número de atruibutos associados às variáveis que em face da amostra considerada poderá não constituir um conjunto suficientemente grande para tornar a análise robusta". Os mesmos autores entendem ainda que o agrupamento de factores tem "tanto de arte como de ciência" e é comum a observância de algumas estatísticas que indiciem a forma como se deverá proceder. As mais conhecidas são as estatística de χ^2, e "*information value" e o weigth of evidence.*

Capítulo IV

4 Modelo de regressão logística (Logit)

Este capítulo descreve o trabalho de modelação empreendido para avaliar o risco de crédito do cliente à luz do modelo de regressão logística. Inicia-se assim com uma apresentação sumária das suas origens. Depois apresenta-se o modelo teórico, modelo logit ou regressão logística binomial (dois nomes para o mesmo modelo). De seguida descreve-se os pressupostos do modelo e as suas estatísticas de avaliação dos diferentes modelos estimados. Por último, serão referidas medidas da qualidade dos ajustamentos como o teste de Hosmer e Lemeshow e a curva de ROC.

4.1 Regressão logística história

A regressão logística surgiu em 1789, com os estudos de crescimento populacional de Malthus. Segundo, Cramer 2002, 40 anos depois, Alphonse Quetelet, astrónomo Belga e, o seu discípulo Pierre- François Verhust (1804-1849), recuperaram a ideia de Malthus para descrever o crescimento populacional em França, Bélgica e Rússia antes de 1833. Apesar de estar encontrada a ideia básica do modelo logístico, só em 1845, Pierre-François Verhust publicou a formulação utilizada nos estudos de crescimento da população a que chamou de curva logística, sendo a expressão matemática a seguinte:

(4.1)

Ainda no séc. XIX, a mesma função foi utilizada para descrever as reacções químicas autocatalíticas, mas esteve esquecido nas neblinas do tempo a maior parte do século e, só foi redescoberto em 1920 por Raymond Pearl, discípulo de Karl Peason, e Lowell Reed que o aplicaram igualmente ao estudo do crescimento da população dos Estados Unidos da América. O primeiro estudo académico abordando a sua aplicação no domínio de *credit scoring* foi publicado em 1980, e desde então tornou-se a técnica estatística de eleição nos desenvolvimentos de modelos de *scoring*.

4.2 Especificação do modelo

Segundo a especificação do modelo clássico de regressão linear múltipla, o comportamento de uma variável dita dependente (também designada por resposta, resultado ou endógena) é uma função de um conjunto de variáveis independentes (também designadas de exploratórias, pré-determinadas ou exógenas). Frequentemente, a variável que se pretende explicar (variável dependente) é de natureza qualitativa, assumindo, um número reduzido de valores, com uma probabilidade diferente associada a cada um destes valores. Por exemplo, nos modelos de *credit scoring* em que a variável dependente (probabilidade de um cliente vir a entrar em situação de incumprimento), é de natureza binomial ou dicotómica, ou seja, pode apenas assumir dois valores (regular, *default*).

(4.2)

Existem vários modelos para explicar . Antes de derivar o modelo de regressão logística vamos introduzir o modelo de probabilidade linear (MPL).

4.2.1 Modelo de Probabilidade Linear

Seja a seguinte especificação:

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \quad (4.3)$$

Sendo o valor esperado de y_i por definição igual a:

$$E(y_i) = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} \quad (4.4)$$

Mas como y_i apenas pode assumir dois valores, o seu valor esperado é também dado por:

$$E(y_i) = 0.(1 - P_i) + 1.P_i = P_i \quad (4.5)$$

Em que P_i é a probabilidade de y_i assumir valor 1

Donde se deduz:

$$P_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} \quad (4.6)$$

Como este modelo exprime a probabilidade P_i como uma função linear das variáveis explicativas, é conhecido como *Modelo de Probabilidade linear.* De notar que P_i, ou seja, a probabilidade de y_i assumir o valor 1 (do cliente ser considerado em situação de *default*) aumenta linearmente com a variação de uma qualquer das variáveis explicativas.

Por outro lado, como a probabilidade deverá situar-se entre zero e um, o modelo de probabilidade linear deverá verificar a restrição:

$$0 \leq P_i \leq 1 \quad (4.7)$$

O que dificilmente acontece já que, a probabilidade cresce linearmente com as variáveis explicativas. De salientar, igualmente, que para além de y_i assumir qualquer valor na recta real, o MPL também não satisfaz as hipóteses de normalidade e homocedasticidade do modelo de regressão clássica.

Verifica-se, com efeito, que os erros assumem apenas dois valores (com probabilidade e , assumindo portanto uma distribuição binomial com média igual a zero e variância igual a a qual não é constante.

Assim, o MPL apresenta vários problemas, o que levou à opção por outras especificações. Entre estas especificações, uma das mais conhecidas[9] é o modelo da regressão logística.

[9] A outra especificação mais conhecida é designada por modelo Probit ou Normit que utiliza a distribuição normal como aproximação.

4.2.2 Derivação do Modelo de Regressão Logística Binomial.

Dado, então, a variável de resposta binária y com probabilidade de sucesso , a regressão logística é um modelo de regressão não linear com a seguinte formulação proposta por Pierre-François Verhulst:

$$P_i = E(y_i / x_i) = \frac{e^{\beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}}{e^{\beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}} \qquad (4.8)$$

Que se pode escrever:

$$P = E(y_i / x_i) = \frac{e^{z_i}}{e^{z_i}} \qquad (4.9)$$

Com $z_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$

Demonstra-se facilmente que:

z

z

Assim no modelo logit P_i é crescente sem nunca assumir valores fora do intervalo [0,1].

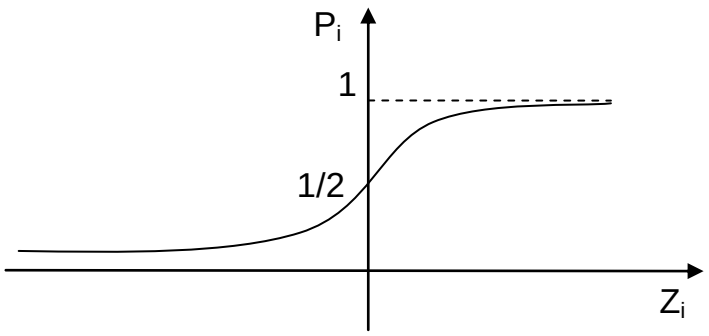


Figura 4.1 – Função logística

Por outro lado, o modelo (4.9) pode ser facilmente linearizado. Com efeito, verifica-se:

$$1-P_i = \frac{1}{1+e^{z_i}} \tag{4.10}$$

E, portanto

$$\frac{P_i}{1-P_i} = e^{z_i}$$

O quociente $\frac{P_i}{1-P_i}$ pode ser interpretado muito simplesmente como o (*odds*), rácio de chances, ou probabilidades. Assim, no caso em estudo, este rácio representa a probabilidade de um cliente ser classificado como *default* sobre a probabilidade do mesmo ser classificado como regular.

Se aplicarmos o logaritmo neperiano à transformação (4.10) e adicionarmos a componente residual, obtemos um modelo de regressão logística linearizado:

$$L_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \tag{4.11}$$

Com:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right)$$
$$z_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

A transformação evidenciada em (4.11) resolve as principais dificuldades do modelo de probabilidade linear. Como refere (Hosmer e Lemeshow 2000), “a importância dessa transformação é que L_i tem muitas propriedades desejáveis do modelo de regressão

linear. O modelo *logit* é linear nos seus parâmetros" tem domínio em $\Re$, dependendo dos valores de X, e, em que $P_i \in (0,1)$, conforme decorre da definição de probabilidade.

4.2.3 Estimação do modelo

Se escrevermos o modelo de regressão logística linearizado, teremos:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \qquad (4.12)$$

Este modelo continua a apresentar erros heterocedásticos (com variância não constante), não se aconselhando a estimação do modelo pelo método dos Mínimos Quadrados Ordinários.

Mas a principal dificuldade reside na possibilidade de L_i assumir valores sem [10]significado. Com efeito, P_i assume, em geral, os valores 1 (quando o acontecimento se verifica) ou 0 (no caso oposto) pelo que P_i assume os valores de Ln( ) e de Ln(0), os quais não têm qualquer significado, tornando impossível a estimação do modelo (4.11).

Por esta razão, o modelo de regressão logística não é, em geral, estimado pelo método dos mínimos quadrados, mas sim pelo de máxima verossimilhança.

Seja então, a função de máxima verosimilhança L:

$$L = \prod_{i=1}^{n} f(y_i) \qquad (4.13)$$

Onde n é o numero de indivíduos (igual ao de observações) e $f(y_i) = P_i^{y_i}(1-P_i)^{1-y_i}$ a função densidade de probabilidade de y_i.

[10] De facto se o problema fosse apenas heterocedasticidade, resolver-se-ia facilmente pela transformação do modelo num modelo de regressão clássica o que passa por multiplicar o modelo pelo inverso do desvio padrão dos erros.

Substituindo (4.12) pela expressão da função de probabilidade de y_i, obtém-se:

$$L = \prod_{i=1}^{n} P_i^{y_i} (1 - P_i)^{1-y_i} \tag{4.14}$$

Substituindo P_i, pela função de distribuição logística vem;

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-X_i \beta}} \right)^{y_i} \left(\frac{e^{-X_i \beta}}{1 + e^{-X_i \beta}} \right)^{1-y_i}$$

Representando de modo mais simplificado fica:

$$L(\beta) = \prod_{i=1}^{n} \lambda(X_i \beta)^{y_i} (1 - \lambda(X_i \beta)) \tag{4.15}$$

Onde X_i é o vector $(1 \times k)$ de observações das k variáveis explicativas do indivíduo i, β é o vector dos k parâmetros a estimar e $\lambda(X_i \beta)$ é a função distribuição da logística.

A maximização desta função é um problema equivalente à maximização do seu logaritmo, já que a função logaritmo é uma função monótona crescente. Para facilitar a obtenção do maximizante, tem-se o logaritmo da função de verosimilhança, ou função log-verosimilhança.

$$l(\beta) = \sum_{i=1}^{n} y_i \ln(\lambda(X_i \beta) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - \lambda(X_i \beta)) \tag{4.16}$$

O estimador de máxima verosimilhança dos k componentes de β corresponde, por definição aos valores desses parâmetros que maximizam l. Para obter este máximo, torna-se necessário calcular a primeira e a segunda derivadas de l, designadas respectivamente por Gradiente G e pela matriz Hessiana H. No Maximo de l, o gradiente tem de ser igual a zero e a matriz Hessiana definida negativa.

Demonstra-se que o Gradiente e a matriz Hessiana são respectivamente dados pelas seguintes expressões[11]:

$$G(\beta) = \sum_{i=1}^{n} y_i X - \sum_{i=1}^{n} \lambda(X_i \beta) X_i \qquad (4.17)$$

$$H(\beta) = -\sum \lambda(X_i \beta)(1 - \lambda(X_i \beta) X_i X_i' \qquad (4.18)$$

A expressão (4.16) e (4.17) não permitem calcular por via analítica a solução de $\hat{\beta}$ para β que garante o máximo de $l(\beta)$, ou seja, a solução, onde $G(\hat{\beta})$=0 (condição necessária) e a matriz Hessiana $H(\hat{\beta})$ é definida negativa (condição suficiente). Não é, portanto, possível encontrar directamente uma solução para este problema que assegure a condição necessária para o máximo de $l(\beta)$. Assim, este problema de maximização é resolvido através do recurso a um algoritmo de optimização.

Demonstra-se que a resolução deste problema reduz-se a iterar através da expressão:

. $$\beta_{h+1} = \beta_h - H(\beta_h)^{-1} G(\beta_h) \qquad (4.19)$$

Onde β_h é o valor β na iteração de h. De notar que quando $\beta_{h+1} = \beta_h$, o processo convergiu e, por outro lado, o gradiente de $G(\beta)$ é nulo, o que garante a verificação da condição necessária para a existência do máximo.
Um dos algoritmos de optimização mais utilizados é o de *Newton-Raphson*. (Amemiya 1985) demonstra que o *log* da função de verosimilhança é globalmente côncavo. Assim o algoritmo de *Newton-Raphson* converge para um único máximo (os estimadores de máxima verosimilhança) independentemente dos valores de inicialização adoptados.

A matriz de variâncias-covariâncias assimptóticas do vector de parâmetro β pode ser estimada através do inverso da matriz Hessiana $-H(\beta_{MV})^{-1}$, avaliada para os estimadores de máxima verosimilhança (MV), β_{MV}. Os estimadores da diagonal principal correspondem às variâncias e os restantes às covariâncias.

[11] A demostração pode encontar-se em Franses e Paag (2001 p.59-60)

4.3 Testes de significância do modelo

Depois de se obter os coeficientes do modelo, coloca-se a questão de avaliar a qualidade da estimação, o que passa por saber até que ponto as variáveis explicativas pertencentes ao modelo são significativas para explicar o comportamento da variável resposta.

Três dos testes mais utilizados para aferir a bondade global do modelo e a significância individual dos parâmetros ou de um conjunto de parâmetros do modelo são o teste de razão de verosimilhança, o teste de Wald e o teste de scores.

4.3.1 Teste de razão de verosimilhança

O teste de razão de verosimilhança (RV) é baseado no mesmo conceito que o teste F para o modelo clássico de regressão linear. O teste F mede o aumento na soma dos quadrados dos resíduos quando as variáveis são retiradas do modelo.
Na regressão logística, o teste de RV é baseado nas diferenças entre os logaritmos da função verosimilhança para os modelos com e sem restrições. Pela teoria de estimação de máxima verosimilhança, sabe-se que os estimadores de máxima verosimilhança maximizam a função log-verosimilhança, pelo que retirar as variáveis resultam geralmente num valor pequeno para a log-verosimilhança, a semelhança do que acontece com o R^2 no modelo de regressão clássica. Isto é, similar ao facto de R^2 nunca aumentar quando algumas variáveis são retiradas da regressão. Com efeito, o teste de razão de verosimilhança avalia se o valor de log-verosimilhança é suficientemente grande para concluir que as variáveis retiradas são importantes para o modelo.

O teste do rácio de máxima verosimilhança baseia-se, portanto no valor obtido pelo rácio:

$$\lambda = \ln \frac{l(\hat{\beta}_R)}{l(\hat{\beta}_U)}$$

O que é por definição igual à diferença:

$$\lambda = l(\hat{\beta}_R) - l(\hat{\beta}_U) \tag{4.20}$$

Onde $l(\hat{\beta}_R)$ é igual ao valor máximo do logaritmo de log-verosimilhança com os $k-1$ parâmetros (todos, excepto a constante) iguais a 0 e $l(\hat{\beta}_U)$ é o valor máximo do logaritmo da função de máxima verosimilhança (sem restrições).

Salienta-se que, quando todos os parâmetros (excepto a constante β_1) são nulos, se verifica, $p_i = \lambda(y_i\beta) = \dfrac{1}{1+e^{-\beta_1}} = p$,

Pelo que $l(\beta_R) = \sum_{I=1}^{n} y_i \ln(p) + \sum_{i=1}^{n} (1-y_i)\ln(1-p)$

E, portanto,

$$l(\hat{\beta}_R) = n(p\ln(p) + (1-p)\ln(1-p)) \tag{4.21}$$

Na hipótese de H_0 ser verdadeira, ou seja, de todos os parâmetros das $k-1$ variáveis explicativas serem nulos, então [12] $RV = -2\lambda$ tem a distribuição do χ^2 com $k-1$ graus de liberdade (igual ao número de restrições).

Supondo $\overline{\beta} = \beta_1, \beta_2, ...\beta_k$, a hipótese a testar é a seguinte.

$$\begin{aligned} H_0 &= \overline{\beta}=0 \\ H_1 &= \exists\overline{\beta} \neq 0 \end{aligned} \tag{4.22}$$

Rejeitando-se a hipótese nula quando *p-value* <0.05, concluindo-se que a informação acerca das variáveis independentes permite-nos realizar previsões estatisticamente válidas.

[12] A multiplicação por 2 é necessária para que a estatística RV tenha uma aproximação à distribuição do qui-quadrado sob a hipótese H_0

4.3.2 Teste de significância dos parâmetros (testes de Wald)

O teste de Wald pode ser obtido, comparando a estimativa de máxima verosimilhança de determinado coeficiente, $\hat{\beta}_j$, com a estimativa do seu erro padrão. Assim, as hipóteses são as seguintes:

$$\begin{aligned} H_0 &= \hat{\beta}_j = \beta^*_j \\ H_1 &= \hat{\beta}_j \neq \beta^*_j \quad (j = 2,...k) \end{aligned} \tag{4.23}$$

E a estatística teste dada pela seguinte expressão:

$$w_j = \frac{\hat{\beta}_j - \beta^*_j}{\sqrt{\text{var}(\hat{\beta}_j)}} \tag{4.24}$$

Onde $\sqrt{\text{var}(\hat{\beta}_j)}$ é o desvio padrão estimado do estimador do parâmetro β_j .A estatística w_j apresenta uma distribuição qui-quadrado com número de graus de liberdade igual ao número de restrições (no caso presente apenas um). Os valores críticos, α_j, para as estimativas dos parâmetros são os níveis para os quais, se o valor do teste de Wald calculado para um determinado β_j for maior que o α_j, se rejeita a hipótese nula para um dado nível significância.

4.3.3 Teste de score (teste de multiplicadores de Lagrange)

Outro teste utilizado para avaliar a significância das variáveis explicativas, é o teste de *scores*, também conhecido como teste de multiplicadores de Lagrange. Este teste pode tornar-se vantajoso comparativamente aos testes anteriormente referidos pelo facto deste não requerer o cálculo da estimativa de máxima verosimilhança para os parâmetros do modelo.

Este teste apenas estima o modelo com restrições e avalia o declive da função log-verosimilhança na restrição. Se a hipótese for verdadeira, o declive (*score)* na restrição deverá ser próximo de zero.

Quando a hipótese nula é verdadeira, o teste de *Wald*, rácio de verosimilhança, e *scores* são assimptoticamente equivalentes. Quando N aumenta, a distribuição amostral dos três converge para a distribuição do qui-quadrado com graus de liberdade igual ao número de restrições testadas.

4.4 Medidas de associação múltipla entre variáveis as independentes e a variável dependente.

Os coeficientes de determinação calculados no modelo de regressão clássica não são aplicáveis no presente caso, visto a variável dependente assumir apenas dois valores. Foram assim desenvolvidas outras fórmulas de cálculo.

4.4.1 Pseudo R^2 (teste de McFadden)

O ρ^2 de McFadden é uma das medidas de associação múltipla entre as variáveis independentes e a dependente mais conhecida na regressão logística. Proposto em McFadden (1974), é uma transformação da razão de verosimilhança na tentativa de se assemelhar ao R^2 da regressão clássica com valores entre 0 e 1:

$$\rho^2 = R^2_{McF} = 1 - \frac{l(\hat{\beta}_R)}{l(\hat{\beta}_U)} \tag{4.25}$$

Onde o significado dos símbolos é o mesmo que em (4.20)

O valor de ℓ^2 está limitado entre (0 e 1), assumindo o valor mínimo, zero, quando $l(\hat{\beta}_R) - l(\hat{\beta}_U)$ Por outro lado, só atinge o valor 1 se a aproximação for sempre perfeita, ou seja $p_i = 1$, quando $y_i = 1$ e $p_i = 0$ quando $y_i = 0$. Apenas, num tal caso $l(\hat{\beta}_R)$ é igual a zero. Entre estes dois limites (0 e 1), o valor de ℓ^2 não tem uma interpretação óbvia, sendo no entanto, valores mais elevados destes coeficientes associados, em geral, a

maior capacidade explicativa do modelo. Segundo (Tabachnick e Fidell 2001) valores entre 0,2 e 0,4 consideram-se satisfatórios.

4.4.2 R² de Cox e Snell

Esta medida baseia-se no logaritmo da função máxima verosimilhança e leva em linha de conta a dimensão da amostra.

$$R_{CS}^2 = 1 - e^{\frac{2}{n} l(\hat{\beta}_R) - l(\hat{\beta}_U)} \quad (4.26)$$

O valor de R_{CS}^2 nunca atinge o valor Maximo 1 e considera-se uma boa aderência valores acima de 0,22.

4.4.3 R² de Nagelkerke

Foi proposto por Nagelkerke, deriva do R_{CS}^2 e, assim, o valor 1 pode ser atingido.

$$R_N^2 = \frac{R_{CS}^2}{R_{\max}^2} \quad (4.27)$$

Em que, $R_{\max}^2$ é o valor máximo de R_{CS}^2 , ou seja, o valor quando, $l(\hat{\beta}_U) = 0$

$$R_{\max}^2 = 1 - e^{\frac{2}{n} l(\hat{\beta}_R)} \quad (4.28)$$

De notar que R_N^2 tem como máximo o valor 1.
Valores de R_N^2 acima dos 0.3 são considerados tradutores de boa qualidade de ajustamento.

Também a este nível não podemos interpretar estes coeficientes do mesmo modo que o R^2 (coeficiente de determinação) no caso de modelos de regressão clássica. Podemos, no entanto associar a valores mais elevados de R^2_{CS} e R^2_N maior capacidade explicativa do modelo.

4.5 Medidas de qualidade do ajustamento

Após a estimação do modelo, o mais adequado é avaliar a qualidade do ajustamento do mesmo. A avaliação da qualidade do ajustamento pode ser feita através dos seguintes testes:

- Testes de Hosmer e Lemeshow,
- Curva de ROC;

4.5.1 Testes de Hosmer e Lemeshow

O teste de Hosmer e Lemeshow é um procedimento habitual para avaliar a qualidade de ajustamento aos dados num modelo de logit. Os seus autores sugerem que o intervalo [0,1] de variação de probabilidade, p_i seja dividido em g intervalos mutuamente exclusivos (aproximadamente 10), comparando-se de seguida as frequências esperadas e as observadas em cada grupo.

A elaboração dos testes consiste nos seguintes passos:

1. Para cada observação estima-se a probabilidade de sucesso;
2. Ordenam-se as probabilidades estimadas por ordem crescente;
3. Agrupam-se os dados de acordo com os decis das probabilidades estimadas;
4. Em cada decil, dividem-se as observações e os valores esperados para o sucesso e insucesso;
5. Calculam-se as estatísticas de teste da à semelhança do cálculo de uma tabela de contingência.

A hipótese a testar é a seguinte:

$$H_0 = o_j = \mathrm{e}_j, \forall_j = 1,...,\mathrm{g}$$
$$H_1 = \exists j : o_j \neq \mathrm{e}_j, j = 1,...,\mathrm{g} \quad (4.29)$$

A estatística de teste sob a hipótese nula é a seguinte:

$$\chi^2_{HL} = \sum_{j=1}^{\mathrm{g}} \frac{(\mathrm{o}_j - \mathrm{e}_j)^2}{e_j\left(1 - \frac{e_j}{n_j}\right)} = \sum_{j=1}^{g} \frac{(\mathrm{o}_j - \overline{\mathrm{p}}_j)^2}{n_j \overline{\mathrm{p}}_j (1 - \overline{\mathrm{p}}_j)} \sim \chi^2_{g-2} \quad (4.30)$$

Onde f- número de decis;

n_j - Número de observações pertencentes ao grupo *j*, verificando-se $n = \sum_{j=1}^{g} n_j$

o_j Frequência observada de sucesso no grupo j, onde $o_j = \sum_{i=1}^{n_j} y_{ij}$ e y_{ij} é a i-ésima observação do grupo j.

e_j - Frequência esperada de sucesso no grupo j, onde $\mathrm{e}_j = n_j \overline{p}_j, \overline{p}_j = \frac{\sum_{i=1}^{n_j} \hat{p}_{ji}}{n_j}$

$\hat{p}_j$ É a probabilidade estimada correspondente à i-ésima observação do grupo j.

Rejeitando-se H_0 quando $\chi^2_{HL} > \chi^2_{g-2,1-\alpha}$, para um nível de significância fixado, α. Análise da estatística χ^2_{HL} fornece uma indicação da qualidade de ajustamento do modelo, e assim, valores grandes desta estatística evidenciam fraca aderência aos dados.

4.5.2 Análise de resíduos

Conforme refere (Chorão 2005) “O principal próposito da análise de resíduos da regressão logística, é identificar as observações para os quais o modelo tem pouca aderência ou observações que exercem mais do que a sua quota-parte de responsabilidade na estimação dos parâmetros do modelo” (pag. 43).

A este próposito, (Chorão 2005) reitira a importância da identificação e posterior remoção das obervações tidas como *outliers*, porém apela ao bom senso e a uma análise critériosa das observações a retirar, pois que, um cliente em situação de *default* é, por si só, um indivíduo atípico. Existem dois tipos de resíduos: O resíduo de Pearson e o resíduo *deviance.*

O resíduo de Pearson é a diferença para cada observação entre o valor observado e a probabilidade estimada dividida pelo desvio-padrão binomial da probabilidade estimada.

$$e_i = \frac{P\ y_i = 1\ - p_i}{\sqrt{p_i \times\ 1 - p_i}} \tag{4.31}$$

Para grandes amostras, o resíduo de Pearson segue uma distribuição normal com disvio-padrão um. Valores absolutos elevados indicam que o modelo não tem aderência à observação em particular.

Normalmente existe nas bases de dados de análise, um conjunto pequeno de observações muito diferentes das restantes. A análise estatística é muito sensível a estas observações, na medida que, uma mudança residual no valor das mesmas provoca uma alteração brusca no valor da estimativa da variável dependente. Estas observações denominam-se *leverage points*, ou observações influêntes e a fórmula para a sua dedução é a que de seguida se apresenta:

Matriz das variáveis explicativas $X_{(n\times p)}$;

Matriz diagonal – V- de dimensão $(n\times n)$, constituida pelo produto entre a Probabilidade estimada e o seu complementar

$$V = diag\ \ p_i(1 - p_i) \tag{4.32}$$

O vector h é então determinado pela seguinte relação

$$h_i = p_i(1-p_i)\ x_i \left[\mathrm{x}^{\mathrm{T}} \mathrm{vx} \right]^{-1} x_i^t \qquad (4.33)$$

Geralmente a estatística de *leverage* assume valores no intervalo (0,1), porém quando a equação do modelo inclui o termo intercepto, poderá assumir valores maiores que 1 ou 1/N.

Valores elevados de h_i indicam grande influência da observação i e um valor igual a 1 significa que o vector dos parâmetros é influnênciado em 100% pela observação em causa.

Numa equação com k variáveis independentes, $\sum_{i=1}^{n} h_i = k+1,$ a média de h_i é dado por $\frac{(k+1)}{N}$ sendo N o número de observações. São de levar em linha de conta para a análise, observações com *leverage* maior que a média.

A distância de Cook é outro indicador utilizado para aferir o impacto da observação i no vector dos parâmetros estimado $\hat{\beta}$. Indica a variação nos resíduos em virtude da eliminação da observação i

$$C_i = \frac{e_i h_{ii}}{k(1-h_{ii})} \qquad (4.34)$$

Onde e_i é o resíduo de Pearson definido anteriormente em (4.31), h_{ii} o *leverage* (4.33) e K o número de parâmetros do modelo.

Dbeta é uma medida estandardizada desta estatística. Valores maiores que 1 merecem uma análise mais cuidada.

4.5.3 Curva ROC

A curva de (*Receiver Operating Characterisctic)*, também conhecido como curva de Lorenz (Henley e McNeil 1982), é bastante utilizada na área médica, para especificar problemas no desempenho de diagnósticos médicos, em que se procura identificar a presença ou ausência de certa doença, com determinada probabilidade de erro. Na área de *credit scoring* é uma técnica bastante útil para avaliar o desempenho de modelos de *scoring.* De uma forma imediata a curva é baseada nos conceitos de sensibilidade e especificidade estatísticas (medida de taxa de classificações correctas) que podem ser obtidas a partir da construção de matrizes de confusão (Johnson e Wichern 2002) obtidas do resultado da classificação dos indivíduos, gerada pelo modelo.

Com o modelo ajustado, a partir de uma amostra de n cliente atribui-se um *score* S para cada indivíduo. Assim o i-ésimo indivíduo será classificado como um *default* se $S_i \leq C_o$, (em que C_o é o *cutoff* para o *score* S_i, previamente determinado), e como regular caso contrário. Para um determinado *cutoff*, é possível determinar a matriz de confusão, como apresentado na tabela seguinte (tabela-4.1)

Tabela 4.1 –Matriz de confusão para duas classes.

		Valores Previstos		Erros	
		Regulares	*Defaults*	Regulares	*Default*
Valores Observados	Regulares	a (TN)	b (FP)		$\frac{b}{a+b} = \beta$
	Defaults	c (FN)	d (TP)	$\frac{c}{c+d} = \alpha$	

Nota: TN- *True negative;* FP – *false positive;* FN- *false negative;* TP – *True positive*

Através da matriz de confusão é possível determinar a percentagem de classificações correctas do modelo ajustado, que são as medidas de especificidade (probabilidade de um cliente ser regular, por aplicação do modelo, sendo-o realmente *score* menor que o c*utoff*) e de sensibilidade (probabilidade de um indivíduo, através do modelo, ser

classificado em *default* quando o é efectivamente – *score* superior ou igual ao *cutoff),* ou seja:

Sensibilidade= $\frac{TP}{TP+FN}$

Especificidade = $\frac{TN}{TN+FP}$

Pode-se igualmente, para um determinado *cutoff* determinar o erro total do modelo de classificação, dado pela seguinte percentagem:

$$\frac{b+c}{a+b+c+d}\times 100\,\% \qquad (4.35)$$

Citando (Chorão 2005), há a realçar nesta matriz vários aspectos importantes associados a problemas de *credit scoring.* Assim,

1. Erro tipo I

   Designado por α (dimensão do teste) ou por risco de crédito, é o rácio de clientes em situação de *default* classificados como sendo regulares. Se uma instituição financeira tem uma taxa α elevada, significa que é demasiado generosa na concessão de crédito estando, portanto, exposta a risco de crédito.
2. Erro tipo II

   Designado por β (complementar da potência de teste) ou por risco comercial, é o rácio de clientes regulares classificados como *defaults.* "Quando, numa instituição financeira, β é elevado por um longo período, haverá perdas nas vendas e concomitantemente, quebras nos lucros. A instituição está exposta ao risco comercial, i.e., ao risco de perda de quota de mercado.
3. *Cutoff*

   α e β estão dependentes do *cutoff* considerando para classificar um cliente em regular ou *default*. Além disso, a matriz de confusão é muitas vezes usada para comparar diferentes modelos de classificação, tendo como hipótese que os dois tipos de erros têm a mesma importância para a instituição.

4.5.3.1 Área abaixo da curva ROC

A área abaixo da curva de ROC, que varia entre 0 e 1, fornece uma medida da capacidade do modelo discriminar entre indivíduos com factor de interesse versus os que não tem factor de interesse. Contudo, quando se considera um teste onde estão presentes duas populações, uma de indivíduos *defaults* (presença de factor de interesse), outra de indivíduos regulares (ausência de factor de interesse), muito raramente se observa uma perfeita separação entre as duas populações. Regra geral, os resultados do teste apresentam uma sobreposição conforme se denota na figura 4.2.

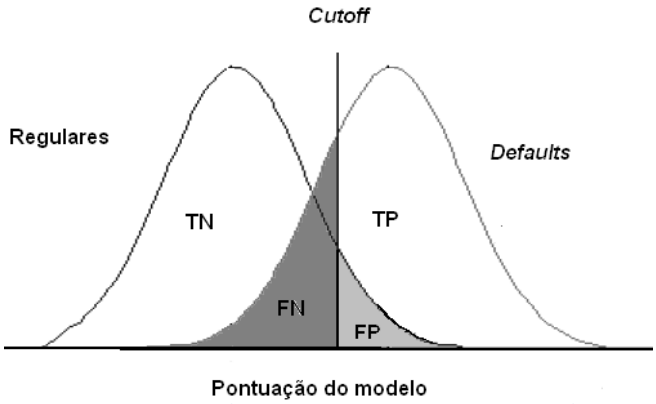


Figura 4.2 - Funções de densidade de duas populações

Para a direita do *cutoff* (teste positivo) identificamos uma área correspondente aos *false positive* (FP) e outra ao *true positive* (TP). Para a esquerda do *cutoff (teste* negativo) identificamos uma área correspondente aos *false negative* (FN) e outra aos *true negative.* Quando menor for a sobreposição das distribuições, menor é a área correspondente aos *false positive.* Assim, de acordo com (A. C. Braga 2000) valores de corte elevados conduzem a um teste pouco sensível e muito específico; por outro lado, valores de *cutoff* baixos conduzem a um teste muito sensível e pouco específico.

Geometricamente, a curva ROC é um gráfico de pares de “x” e “y” (que correspondem, a 1 - especificidade e à sensibilidade, respectivamente) num plano designado por plano ROC unitário. Deste modo, no eixo das ordenadas está representada a sensibilidade do modelo, isto é, quão bom é o modelo é em prever os *true positives (defaults*) sendo as suas coordenadas calculadas a partir de:

$$Y=\frac{TP}{TP+FN} \tag{4.36}$$

No eixo das abcissas encontra-se o complementar da especificidade, isto é, a capacidade do modelo não errar na identificação dos *true negatives* (regulares). Em geral, um aumento da sensibilidade implica um decréscimo na especificidade e vice-versa. As do eixo das abcissas são calculadas a partir de:

$$X=1-\frac{TN}{FP+TN} \tag{4.37}$$

A relação anterior encontra-se ilustrada na figura 4.3 onde se alude, igualmente, aos três tipos de modelo que a figura permite inferir.

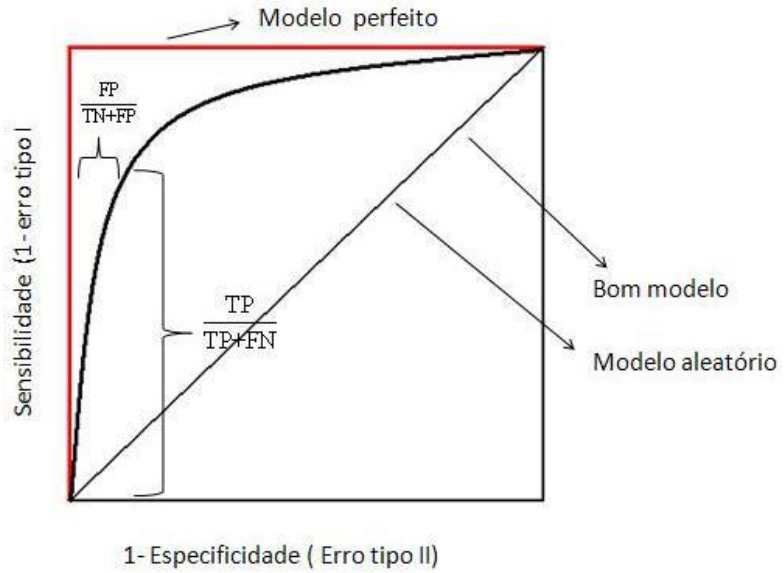


Figura 4.3 - Curva de ROC, com apresentação das coordenadas para um dado cutoff

Quando maior a sensibilidade para valores elevados da especificidade (ou seja, valores elevados do eixo dos y's e valores baixos dos x's) melhor o modelo estimado. Neste sentido, uma medida numérica da precisão pode ser obtida pela área da curva, em que o valor 1 significa um modelo perfeito, enquanto uma área em redor de 0,5 indica uma fraca capacidade de aderência aos dados (modelo aleatório). Ou dito de outra forma, a área delimitada pela curva mede a discriminação, isto é, a capacidade preditiva do modelo classificar correctamente os indivíduos em *defaults* e os individuos regulares

Capítulo V

5 Redes Neuronais Artificiais

As redes neuronais artificiais são modelos que surgiram originalmente na década de 1940 como tentativa de reprodução do funcionamento do cérebro humano, sendo o complexo sistema de neurónios biológicos a sua principal fonte de inspiração. A desmistificação deste conceito torna-se assim inevitável e se de facto as redes neuronais estão relacionadas com o cérebro biológico o seu estudo e desenvolvimento envolve para além da neuro-biologia, outras áreas do conhecimento tais como a matemática, a electrónica e a cibernética[13].

Os métodos de neuro-computação são estreitamente baseados num modelo artificial do cérebro como uma rede de elementos de processamento simples conectados entre si, correspondendo aos neurónios biológicos, mas cuja actuação colectiva lhes confere grande capacidade de processamento possuindo estes sistemas como principal vantagem o facto de poderem aprender e adaptar-se a alterações ambientais (Cloete 2000)

Os modelos neuronais têm tido inúmeras aplicações nas mais diversas áreas, desde as telecomunicações ao mercado imobiliário, das despesas militares ao turismo (Shachmurove 2002) e (Law e Pine 2004) desde a robótica à visão (Kröse e Smagt 1996), das relações internacionais (Beck, King e Zeng 2000) às questões de política interna (Eisinga, Franses e Dijk 1997). Na área financeira vários problemas têm sido abordados recorrendo às redes neuronais, como a análise do risco de crédito (Nargundkar e Priestley 2004), a previsão da insolvência de empresas (Neves e Vieira 2004), a modelização da inflação (McNelis 2005), a modelização das taxas de câmbio (Zhang e Lin 2002) , o rating de obrigações, a previsão da volatilidade das opções (McNelis 2005), a previsão das rendibilidades de acções (Thawornwong e Enke 2004) (Zhang e Lin 2002) a previsão de índices e tendências de mercados accionistas.

[13] **Cibernética** é uma teoria da comunicação e controlo do feedback de regulação. O termo cibernético advém do grego Κυβερνήτης (significando condutor, governador, piloto). A cibernética é a disciplina que estuda a comunicação e o controlo nos seres humanos e nas máquinas construídas pelos humanos (http://pt.wikipedia.org/wiki/Cibern%C3%A9tica)

5.1 Inspiração Biológica: O Cérebro Humano

Grande parte da investigação em Redes Neuronais artificiais (RNAs) foi inspirada e influenciada pelo sistema nervoso do ser humano. Muitos investigadores acreditam que as RNAs oferecem a aproximação mais promissora para a construção de verdadeiros sistemas inteligentes, tendo capacidade para ultrapassar a explosão combinatorial associada à computação simbólica baseada em arquitetura de Von Neumann [14]

O sistema nervoso central fornece uma forte base de sustentação a esta tese. O cérebro é uma estrutura altamente complexa, não linear e paralela. Possui uma capacidade de organizar os seus constituintes, conhecidos por neurónios de modo a executarem certas tarefas complexas (e.g processamento em paralelo da informação, a memória associativa e a capacidade para classificar e generalizar conceitos), de uma forma inatingível pelo computador mais potente até hoje concebido.

Apesar dos grandes avanços científicos, o conhecimento do modo como o cérebro humano funciona está longe de estar completo. No entanto, alguns factos importantes já são conhecidos. Quando alguém nasce, o seu cérebro apresenta se já com uma estrutura fortemente conexionista, com capacidade de aprender através da experiência. Este conhecimento evolui através do tempo, apresentando-se um desenvolvimento mais acentuado nos primeiros dois anos de vida. Estima-se que o sistema nervoso humano possui aproximadamente 86 bilhões neurónios ligado uns aos outros através de sinapses, e juntos formam uma grande rede, chamada rede neuronal. (Kohonen 2001) refere que e as ligações existentes entre eles – os axónios – possuem um comprimento tal no seu conjunto que se fossem esticados daria para fazer duas vezes a viagem de ida e volta da Terra à Lua.

(Damásio 1995) escreve que nos “neurónios se identificam três componentes importantes: um corpo celular; uma fibra principal de saída, o axónio; e fibras de entrada ou dendrites. Os pontos nos quais os axónios estabelecem contacto com as dendrites de outros neurónios designam-se por sinapses” (p.65).

Portanto, uma rede neuronal consiste essencialmente num conjunto de unidades de processamento simples (neurónios) que comunicam entre si enviando sinais através de um número elevado de conexões. Em termos biológicos, se a informação acumulada

[14] John Von Neumann (1903 -1957) , matemático húngaro-americano que teve uma grande contribuição na definição da arquitectura de máquinas sequenciais, onde um programa é armazenado na mesma memória de dados que o programa utiliza. Hoje em dia quase todos os computadores são do tipo Von Neumann.

no corpo celular de um determinado neurónio atingir certo limite, o neurónio “dispara”, transmitindo um sinal electroquímico a um neurónio adjacente, através de um canal emissor, o axónio. A extremidade do axónio é composta por ramificações (as sinapses) que por sua vez estão ligadas à estrutura do neurónio receptor através de outras ramificações, as dendrites. (ver figura 5.1)

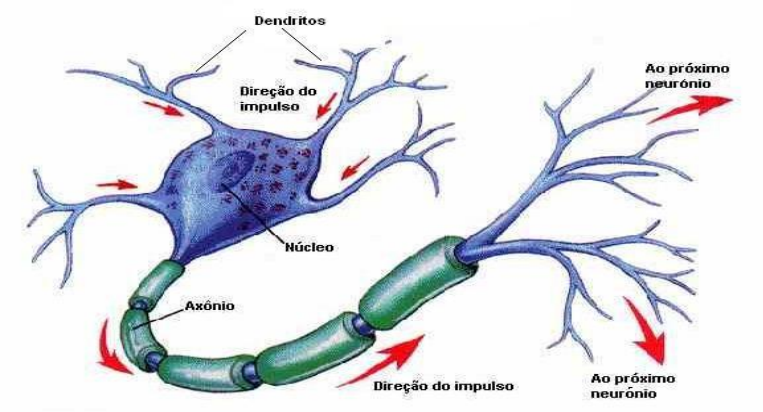


Figura 5.1 -Diagrama de um neurónio

Um único neurónio pode estar ligado centenas ou mesmo a dezena de milhares de neurónios. Num cérebro existem estruturas anatómicas de pequena, média e alta complexidade com diferentes funções, sendo possíveis parcerias. (Cortez e Neves 2000), escreve que os neurónios tendem a agrupar-se em camadas, existindo três principais tipos de conexões: divergente onde o neurónio pode ·estar ligado a vários neurónios via uma arborização do axónio; convergentes, onde vários neurónios podem estar conectados a um único neurónio e encadeadas ou cíclicas, as quais podem envolver vários neurónios e formarem ciclos (ver figura -5.2)

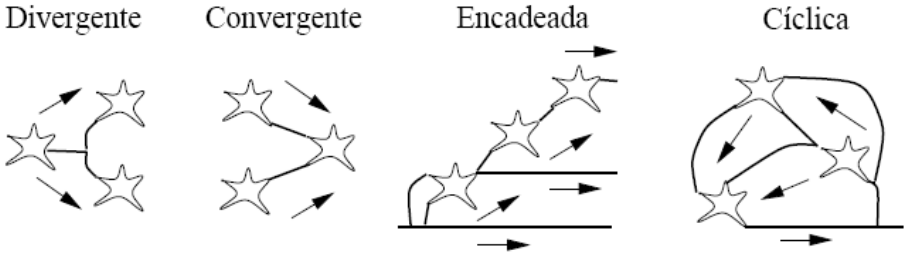


Figura 5.2 - Os diferentes tipos de conexões

5.2 Os componentes de uma Rede Neuronal Artificial

Apesar dos esforços em reproduzir o funcionamento do cérebro humano, tudo o que se conseguiu foi uma aproximação elementar. (Roisenberg e Vieira, Redes Neuronais Artificiais: Um Breve Tutorial s.d.)

Como escreve (Bação 2005), tal como o processo electroquímico de comunicação entre neurónios biológicos, as redes neurónios artificiais[15], também "consistem em neurónios e conexões entre eles. Os neurónios (ou nós) transportam informação de entrada (*input*) e passam a outros neurónios através das suas conexões de saída (*output*). Nas redes neuronais artificiais estas conexões são designadas por pesos ou ponderações (*Weights)*. A informação "*elétrica*" é simulada com valores numéricos específicos armazenados nestes pesos. Através da alteração dos valores dos pesos simulamos a alteração na estrutura de conexão".

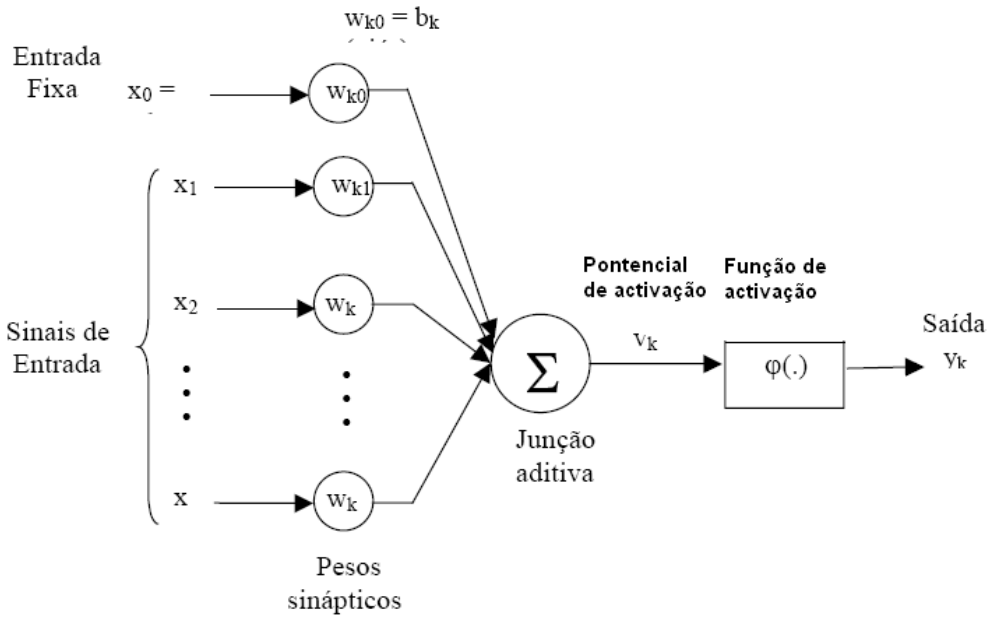


Figura -5.3 -Conceptualização gráfica de um neurónio artificial

Como descreve a figura 5.3, um neurónio artificial é semelhante à célula neuronal biológica, e funciona de forma semelhante. A informação é enviada para o neurónio com base nos pesos de recepção da camada de entrada (*input*). Este *input* é processado por

[15] O termo artificial deriva, obviamente, do facto de estas redes serem implementadas em programas computacionais, capazes de processar o grande número de cálculos necessáriosdurante o processo de aprendizagem.

uma função de combinação que "soma" o valor de todos os *inputs* ponderados recebidos. O valor resultante é comparado com um determinado valor limiar pela função de activação do neurónio. Se o *input* excede o valor limiar, o neurónio será activado e enviará um *output* pelos seus pesos de envio para todos os neurónios a ele conectados e assim sucessivamente, de contrário o neurónio será inibido.

Assim, vista como uma máquina adaptativa, uma rede neuronal é segundo Haykin citado em (Cortez e Neves 2000) "Um processador eminentemente paralelo, composto por simples unidades de processamento, que possui uma propensão natural para armazenar conhecimento empírico e torná-lo acessível ao utilizador".

Assemelha-se ao comportamento do cérebro em dois aspectos:

- O conhecimento é adquirido a partir de um ambiente, através do processo de aprendizagem.

- O conhecimento é armazenado nas conexões, também designadas por ligações ou sinapses entre neurónios.

Durante o processo de aprendizagem, dado por um algoritmo de aprendizagem ou de treino, a força (ou peso) das conexões é ajustada de forma a atingir um desejado objectivo ou estado de conhecimento da rede. Embora seja esta a forma tradicional de construir uma RNAs também é possível modificar a sua própria estrutura interna (ou topologia), à semelhança do que se passa no cérebro, onde neurónios podem morrer e novas sinapses (e mesmo neurónios) se podem desenvolver.
Resumindo, o modelo de neurónio artificial apresentado na figura 5.3 é composto por três elementos básicos:

1. Um conjunto de sinapses (ou ligações conectadas), cada uma as quais caracterizada por um peso, que tem um efeito excitatório para valores positivos e inibitórios para valores negativos. Assim, o sinal ou estímulo
   do input da sinapse ligada ao neurónio é multiplicado pelo peso sináptico correspondente . Pode ainda existir uma ligação extra, denominada de "*bias"*

cuja entrada toma o valor +1, que estabelece uma certa tendência ou inclinação no processo computacional, isto é, adiciona uma constante para que se estabeleçam as correctas condições operacionais para o nodo.

2. Um totalizador para acumular os sinais de entrada. Frequentemente é utilizada a função adição ponderando todos os inputs numa combinação linear.

3. Uma função de activação (f) para restringir a amplitude do neurónio, de saída. A função de activação é também referida como função esmagadora ou ainda restritiva, já que restringe (limita) a amplitude do sinal de saída a um valor finito.

Citando (Haykin 1999), a razão pela qual as redes neuronais artificiais gozam actualmente de grande popularidade alicerça-se em dois aspectos fundamentais: por um lado numa topologia que premeia o paralelismo, e por outro lado, na sua capacidade de aprendizagem e generalização; isto é, conseguir responder adequadamente a novas situações com base em experiências passadas. São estas duas características que tornam possível a resolução de problemas, que de outra forma seriam intratáveis. Isto não quer dizer que as RNAs sejam caixas mágicas que consigam por si dar resposta a qualquer problema. Pelo contrário, precisam não raras vezes de ser integrados com outros sistemas ou paradigmas.

As redes neuronais apresentam ainda, segundo (Haykin 1999), características únicas, que não se encontram em outras ferramentas ou técnicas:

- Capacidade de aprendizagem e generalização, isto é, consegue descrever o todo a partir algumas partes, constituindo-se como formas eficientes de aprendizagem e armazenamento de conhecimento;
- Não linearidade, atendo a que muitos problemas reais são de natureza não linear;
- Adaptabilidade: podendo adaptar a sua topologia de acordo com mudanças do ambiente;
- Robustez e tolerância a falhas: permite processar o ruído ou informação incompleta de forma eficiente, assim como sendo capazes manter o seu desempenho quando há desactivação de algumas das suas conexões e/ou nodos. O que quer dizer que se uma rede neuronal for treinada para um problema

específico, será capaz de atingir bons resultados, mesmo que o problema não seja exactamente igual àquela que foi utilizado durante o treino.

- Flexibilidade, isto é, tem um grande domínio de aplicabilidade.
- Capacidade de processamento paralelo, permitindo que tarefas complexas sejam realizadas num curto espaço de tempo.

5.3 Redes Neuronais Artificiais: História

A curiosidade sobre o cérebro humano e os processos cognitivos e de coordenação há já muito tempo que intrigam o Homem. As primeiras tentativas de explicação de alguns aspectos teóricos, segundo (Kohonen 2001) foram encetadas por filósofos gregos como Aristóteles (384-322 AC), tendo já os filósofos empíricos do séc. XVI algumas visões do sistema nervoso, de onde se destaca a de Descartes (1596-1650). O conhecimento existente hoje sobre o funcionamento do cérebro é o resultado da investigação feita nos últimos 100 anos

Ramón y Cajal em 1894 foi o primeiro a propor uma teoria para o funcionamento do cérebro em termos de unidades constituintes a que denominou de neurónios. Contudo, a tentativa de reprodução do funcionamento do cérebro humano data do início da década de 1940, com o trabalho pioneiro de McCulloch e Pitts (Haykin 1999). Warren McCulloch foi um psiquiatra e neuroanatomista que dedicou 20 anos de investigação na tentativa de reproduzir um evento no sistema nervoso. Por sua vez, Walter Pitts, um prodígio matemático, juntou-se a McCulloch em 1942, tendo ambos, publicado em 1943, "*A logical calculus of the ideas immanent in nervous activity"*, No Artigo os autores descrevem um cálculo lógico das redes neuronais que sumariava os estudos da neurofisiologia e da matemática lógica. Defendiam ainda que o modelo formal do neurónio por eles desenvolvido seguia uma lei "*tudo ou nada*". Provaram que uma rede neuronal constituída por um número suficiente de neurónios e com conexões sinápticas ajustadas apropriadamente e operando de forma síncrona era capaz de processar qualquer função.

Em 1948, foi publicado o famoso livro *Cybernetics* de Winer, descrevendo alguns conceitos importantes sobre o controle, a comunicação e o processamento estatístico de sinais. A segunda edição do livro foi publicada em 1961, adicionando novos conceitos sobre aprendizagem e *Self- Organizing Maps*. No capítulo 2 de ambos os livros Winer, parece compreender o significado físico de mecânica estatística no contexto deste

assunto, mas foi Hopfield, (mais de 30 anos depois) quem conseguiu estabelecer a ligação entre a mecânica estatística e os sinais de aprendizagem.

O próximo desenvolvimento significativo das redes neuronais veio em 1949, com a publicação do livro de Donald Hebb "*The Organization of Behavior*", em que acentuava a ideia de que os parâmetros do modelo do neurónio de McCulloch-Pitts pudessem se auto-ajustar. Estes primeiros estudos das redes neuronais biológicas formaram os fundamentos do que se tornou conhecido como redes neuronais artificiais (RNAs).

Durante a metade da década de 1950 e início de 1960, uma classe de investigadores chamados de "*learning machines*" liderada por Frank Rosenblatt, causaram grande excitação entre pesquisadores da teoria de reconhecimento de padrões, principalmente pela apresentação do livro *"Principles of Neurodynamics".* Nele o autor fornece várias ideias a respeito dos perceptrões, demonstrando que se adicionarem sinapses ajustáveis, as redes neuronais poderiam ser treinadas para classificar certos tipos de padrões. O perceptrão é capaz de classificar entre classes que são linearmente separáveis, tendo sido utilizado para reconhecer caracteres. (Chorão 2005) refere que a característica mais importante do perceptrão é "a apresentação de um algoritmo de aprendizagem capaz de adaptar os pesos internos do neurónio de maneira que seja capaz de resolver o problema da separabilidade linear das classes". O êxito conseguido por esta abordagem fez com que muitos considerassem Rosenblatt como o verdadeiro pai da inteligencia artificial..

Em 1960, Widrow e Hoff introduziram o algoritmo "*Least Mean Square*" (LMS), conhecido como mínimos quadrados, que usaram para formular o Adaline (elemento linear adaptativo). A principal diferença entre o perceptrão apresentado por Rosemblatt, e o Adaline de Widrow situa-se no procedimento de treino. Widrow e seus estudantes propuseram uma das primeiras redes neuronais com camadas capazes de ser treinadas com múltiplos elementos adaptativos, que foi chamada de Madaline (Haykin 1999). Após a apresentação do perceptrão acreditava-se que as redes neuronais (*perceptrons*) poderiam resolver qualquer problema. Contudo, após estes espectaculares desenvolvimentos, a área das redes neuronais conheceu uma grande crise com a publicação do trabalho de Marvin Minsky e Seymour Papert, em 1969 sobre o "*Perceptrons".* Nele chamaram a atenção para algumas tarefas que o perceptrão com

apenas uma camada intermédia, era incapaz de aprender padrões não linearmente separáveis, (o famoso problema do Xor/Ou – exclusivo). Rosenblatt propôs como solução aumentar o número de camadas, mas, apesar de toda a sua visão e perspicácia neste campo, não logrou desenvolver um método de aprendizagem eficaz para estas redes neuronais mais avançadas. Após a publicação do livro de Minsky e Papert, sobre as limitações dos perceptrões, e também, por não haver suporte financeiro para conduzir projectos nesta área, as pesquisas em redes neuronais ficaram esquecidas pelo menos até o início de 1980.

E 1974 aconteceu um facto que viria, mais tarde a proporcionar o renascimento do interesse geral pelas potencialidades das redes neuronais, foi quando Paul Werbos lançou as bases do algoritmo de retro-propagação ("*Backpropagation*"), Porém as potencialidades deste método tardaram a ser reconhecida (Gorni 1994).

Em 1982 John Hopfield publicou com um estudo que chamava atenção para as propriedades associativas de uma classe de redes neuronais que apresentava fluxos de dados multidirecional e comportamento dinâmico, Primeiramente ele demonstrou que a rede possuía estados estáveis e, posteriormente, que tais estados poderiam ser criados alterando-se os pesos das conexões entre os neurónios. No entanto, os primeiros resultados que levaram a retoma do desenvolvimento das redes neuronais só foram publicados em 1986 e 1987, através dos trabalhos do grupo PDP (*Paralled and distributed Procesing*), onde ficou consagrada a técnica de treino por *backpropagation*. Estava então reunidas as condições para o desenvolvimento das redes neuronais. Em 1982, Kohonen (1982) publicou um artigo no qual utilizava mapas auto-organizáveis (SOM) como uma estrutura bi-dimensional, que difere em alguns aspectos do primeiro trabalho de Willshaw e von der Malsburg que também usaram aprendizagem competitiva. Em 1988, Broomhead e Lowe descreveram um procedimento para o projecto de uma rede neuronal (*feedforward*) usando função de base radial, conhecida na literatura como "*radial basis function*" (RBF), que proporcionou um modelo de aprendizagem alternativo ao perceptrão de multiplas camadas. No início dos anos 90, Vapnik e seus colaboradores apresentaram uma poderosa classe de redes neuronais supervisionadas, designadas de *Support Vector Machines*, para a regressão e o reconhecimento de padrões.

Hoje em dia procuram-se não só redes mais eficientes como também melhores algoritmos de treino (Sarle, Neural network 1999). Por outro lado, espera-se que a aplicação de RNAs a outras áreas do conhecimento se generalize, seja à Medicina, à Economia, ao Processamento de Sinal, à Robótica, ou aos Sistemas Periciais, para além da Estatística.

5.4 Tipos de Redes Neuronais Artificiais

As redes neurais artificiais diferenciam-se pela sua arquitectura e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizagem. A arquitectura de uma rede neural restringe o tipo de problema no qual a rede poderá ser utilizada, e é definida pelo número de camadas (camada única ou múltiplas camadas), pelo número de nós em cada camada, pelo tipo de conexão entre os nós e pela sua topologia (Haykin 1999).

Hoje em dia existem milhares de diferentes tipos de redes neuronais, cada uma com as suas próprias potencialidades e caracteristicas. No entanto, a grande distinção é feita entre redes *feedback*- também designadas recorrentes e *feedforward* (alimentação para a frente).

Uma rede neuronal artificial *feedforward* pode ser organizada por camadas, pois não existem ciclos, dado que as conexões são sempre unidireccionais (convergentes ou divergentes) não existindo realimentação. Na sua forma mais simples, uma rede é composta por uma camada de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída. De referir que a camada de entrada não é contabilizada como camada numa RNA dada o facto de nesta não se efectuarem quaisquer formas de cálculo. A segunda classe de redes *feedforward* distingue-se pelo facto de possuir uma ou mais camadas intermédias, cujos nodos são designados por nodos intermédios tendo como função intervir de forma útil entre a entrada e a saída da rede. Ao se acrescentar camadas intermédias está-se a aumentar a capacidade da rede em modelar funções de maior complexidade, uma particularidade bastante útil quando o número de nodos na camada de entrada é elevado. Por outro lado, este aumento também transporta um senão, uma vez que o tempo de aprendizagem aumenta de forma exponencial.

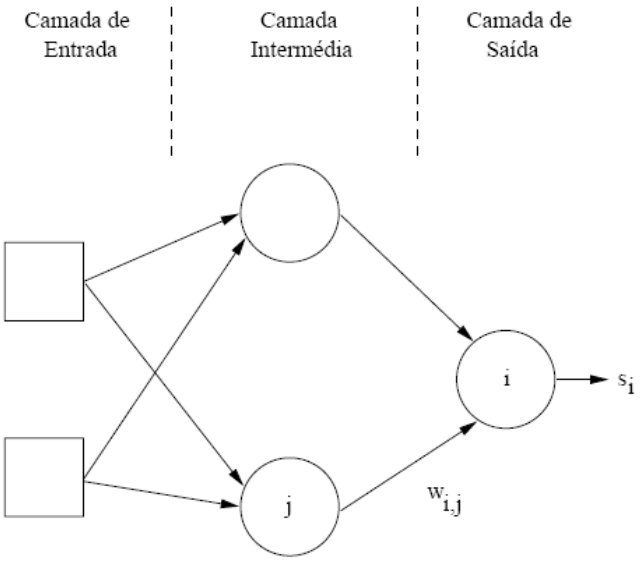


Figura 5.4 - Arquitetura de rede f*eedforward*

Como se poderá observar na figura 5.5, numa rede *feedback*, as conexões podem ser feitas entre quaisquer nodos. A recorrência existe em sistemas dinâmicos quando uma saída de um elemento influência de algum modo a entrada para esse mesmo elemento, criando-se assim um ou mais circuitos. Assim que uma ou mais conexões cíclicas são incluídas numa rede, estas passam a ter um comportamento não linear, de natureza espacial e/ou temporal, que podem ser utilizadas para modelar novas funções cognitivas, tais como as de memória associativa e/ou temporal (Bose e Liang 1996). Ao conter ciclos, as saídas não estão dependentes exclusivamente das ligações entre nodos, mas também de uma dimensão temporal; i.e., está-se na presença de uma cálculo recursivo, que obedecerá naturalmente a uma certa condição de paragem, com a última iteração a ser dada como a saída para o nodo (Riedmiller e Braun. 1993).

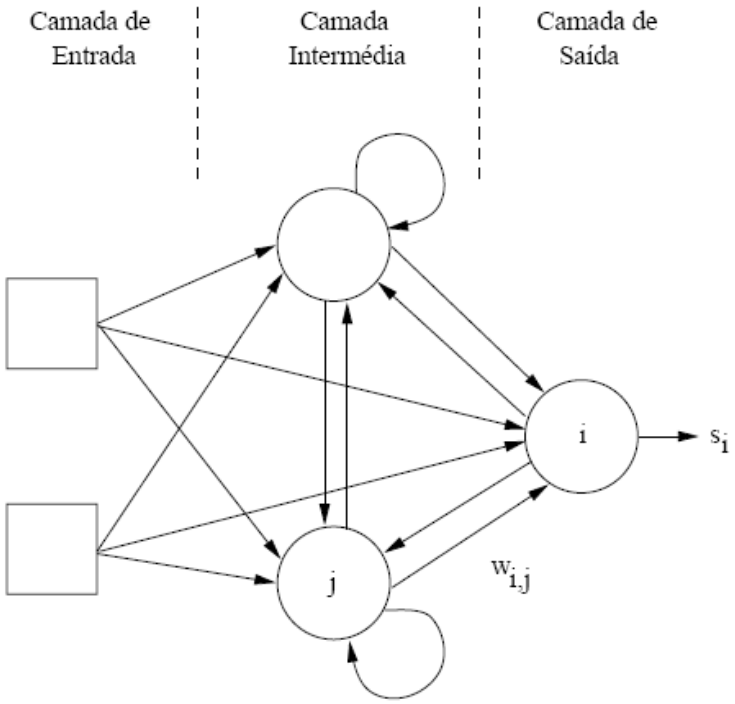


Figura 5.5 - Arquitetura de redes neuronais *feedback*

5.5 Tipos de aprendizagem

Como já foi referido uma das propriedades mais importantes de uma rede neural artificial é a capacidade de aprender a partir da interacção com o meio ambiente e fazer inferências do que aprenderam.

A utilização de redes neuronais num qualquer problema passa primeiramente pela fase de aprendizagem que se dá quando a rede neuronal consegue extrair padrões de informação no subconjunto de treino, criando assim uma representação própria. Segundo (Braga, Carvalho e Ludemir 2000) “a etapa de aprendizagem consiste num processo interactivo de ajustamento dos parâmetros da rede, os pesos das conexões entre as unidades de processamento, que guardam, no final do processo, o conhecimento que a rede adquiriu do ambiente em que se encontra a operar.” (pag. 72)

Enunciando (Haykin 1999) a aprendizagem é um processo pela qual os parâmetros de uma rede neuronal são ajustados através de um processo de estimulação

do meio ambiente no qual a rede está inserida, sendo o tipo de aprendizagem determinado pela maneira como ocorrem os ajustamentos nos parâmetros. Portanto, o objectivo do treino/aprendizagem consiste em atribuir valores apropriados aos pesos sinápticos de modo a produzir o conjunto de saídas desejadas ou ao menos consistentes com um intervalo de erro estabelecido. Desta forma, o processo de aprendizagem consite na busca de um espaço de pesos pela aplicação de alguma regra que defina esta aprendizagem.

Existem três paradigmas básicos para adaptar os parâmetros do sistema: aprendizagem por reforço, aprendizagem supervisionada, e aprendizagem não superviosionada.

5.5.1 Aprendizagem por reforço

Neste tipo de aprendizagem conta-se com a presença de especialistas acerca do universo de discurso, embora a resposta correcta não seja apresentada à rede; i.e., apenas se fornece uma indicação sobre se a resposta da rede está correcta ou errada, tendo a rede de usar essa informação para melhorar o seu desempenho. Em princípio, um prémio dado em termos do reforço dos pesos das conexões que contribuem para uma resposta correcta e, uma penalidade para a situação em contrário.

5.5.2 Aprendizagem Supervisionada

Nas redes neuronais, aprendizagem supervisionada tornou-se a designação do processo de ajustamento de um sistema para que produza um determinado *output*, designado por este motivo de alvo (*target*), como resposta a determinados inputs, sendo a relação funcional existente entre as variáveis independentes e dependente normalmente conhecida e designando-se por treino o processo pelo qual o sistema “aprende” esta relação. Desta forma, a rede pode ajustar os parâmetros de forma a encontrar a solução que melhor adequabilidade registe entre o seu *output* e os seus valores correctos observados.

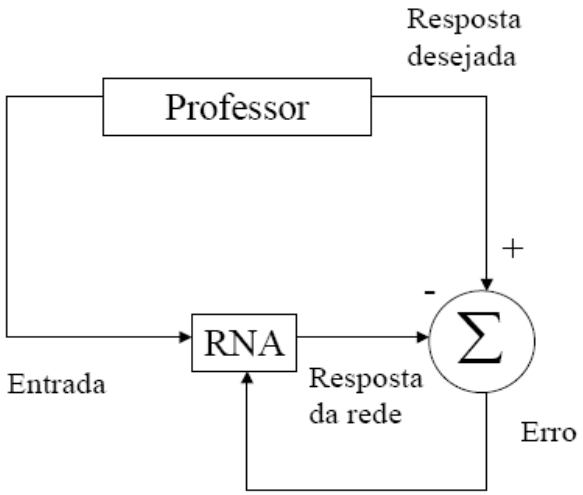


Figura 5.6 - Diagrama do ciclo de aprendizagem supervisionada

Fonte: Adaptado a (Haykin 1999)

Por vezes este tipo de rede refere-se como fazendo uso de um "*professor*" externo que indique ao sistema o correcto resultado para cada padrão de *input*. Podendo este "professor" ser um humano, que especifica a correcta classe para cada padrão de entrada, ou um sistema físico cujo comportamento se pretenda modelar. A cada interacção efectuada a rede neuronal compara a resposta desejada com o valor de saída da rede, originando um erro. O erro resultante é utilizado para de alguma forma ajustar os pesos da rede. A soma dos erros quadráticos de todas as saídas é normalmente utilizada como medida de desempenho da rede. Uma das vantagens da aprendizagem supervisionada é a de que o seu modelo é bem definido, apontando-se como principais críticas o artificialismo, a limitação do modelo de aprendizagem e a necessidade de professor (Reed e MarsII 1999)

5.5.3 Aprendizagem não-supervisionada

Um dos inconvenientes do treino supervisionado é a necessidade de "professor" Suponhamos agora que também não conhecemos *a priori* o número nem as classes envolvidas. Como lidar com situações destas? Para fazer face a estas questões nasceu a necessidade de desenvolver uma aprendizagem e classificação não-supervisionada.

Neste tipo de aprendizagem os dados de treino não se encontram legendados e não existem alvos a atingir, em vez disso o sistema adapta-se às suas idiossincrasias de acordo com as características que possuem implicitamente. Mesmo não conhecendo as classes envolvidas, se as amostras em análise caírem num número finito de categorias, digamos, com base nas suas relações de similaridade, então podemos estar perante um problema de classificação não supervisionada sendo assim necessário recorrer a métodos de agrupamento (clustering) (Kohonen 2001) Este tipo de aprendizagem apresenta ainda a vantagem adicional de poder ser utilizada mais abrangentemente em virtude dos dados não legendados se encontrarem frequentemente em maior disponibilidade que os dados classificados (Reed e MarsII 1999). Se uma rede tiver a habilidade de descobrir clusters com similaridade de padrões sem supervisão, i.e, sem possuir informação sobre o *target*, e a afectar neurónios a esses clusters, qualquer que seja o processo utilizado, diz-se que a rede, além de não ser supervisionada, possui capacidade de auto-organização (Gurney 1997). Um tipo de redes deste tipo e que tem sido muito bem sucedida na resolução e modelação de vários problemas são os *Self-Organizing Maps* (SOM).

5.6 Redes Multi *Layer Perceptron* (multicamadas).

5.6.1 Perceptron de uma única camada.

De acordo com o que se escreveu em 5.3. Redes Neuronais História, pag. 77, foi Rosembalt em 1958, quem primeiramente propôs o *perceptron*, como o primeiro modelo de aprendizagem supervisionada. O *perceptron* é a forma mais simples de uma rede neuronal, que só aceita valores binários (0 e 1) como *input* e como *output*. É utilizada para a classificação de padrões ditos linearmente separáveis, isto é, padrões que se encontram em lados opostos de um hiperplano. Este modelo era composto, basicamente, por um único neurónio com pesos sinápticos ajustáveis e o termo "*bias*". O algoritmo utilizado para ajustar os parâmetros livres desta rede neuronal apareceu primeiro num processo de aprendizagem desenvolvido por Rosembalt para o seu modelo cerebral de *perceptron.* De facto, Rosembalt provou que se os padrões (vectores) utilizados para treinar o *perceptron* são retirados de duas classes linearmente separáveis, então o

algoritmo de *perceptron* converge e posiciona a superfície de decisão na forma de um hiperplano entre as duas classes (vide figura 5.7).

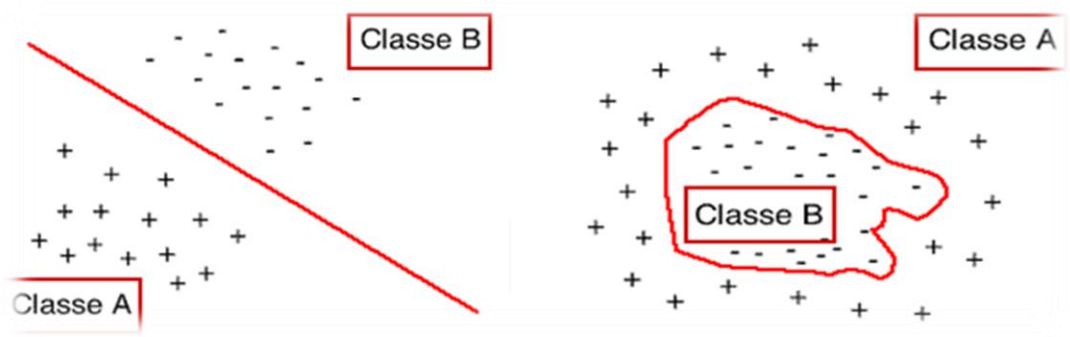


Figura 5.7- Classes linearmente separáve is (à esquerda) e classes não linearmente separáveis (à direita)

Todavia, Minsky e Papert, ao analisarem matematicamente o perceptrão concluíram que este não obtinha soluções para problemas que não fossem linearmente separáveis. Para isso seria necessário a introdução de mais um neurónio na camada intermédia e de uma ou mais camadas intermédias de modo a poder implementar superfícies de decisão mais complexas. A característica principal da camada escondida é que seus elementos se organizam de tal forma que cada elemento aprenda a reconhecer características diferentes do espaço de entrada, assim, o algoritmo de treino deve decidir que características devem ser extraídas do conjunto de treino. Ademais, o algoritmo de minimização do erro, também conhecido como regra delta, apresentado por Windrow e Hoff, não se aplicava às camadas intermédias.

5.6.2 Arquitectura de redes multicamdas (MLP)

Para problemas não linearmente separáveis seria, necessário recorrer a uma combinação de hiperplanos em que se deveria, segundo (Neto 1997) “dotar a rede com mais de um neurónio na camada intermédia, e/ou mais de uma camada. Contudo, na época ainda não se conhecia nenhum algoritmo de aprendizagem capaz de treinar redes com mais de uma camada de neurónios, gerando-se, assim, um grande pessimismo em relação ao futuro da área das redes neuronais”.

Foi somente em 1974 que Paul Werbos, descobriu o algoritmo enquanto desenvolvia a sua tese de doutoramento em Estatística, o qual apelidou de “Algoritmo de realimentação dinâmica”. Parker, em 1982 redescobriu o algoritmo e denominou-o de “Algoritmo de aprendizagem lógico”. Todavia, como referido anteriormente, foi o com o trabalho de Rumelhart Hilton e Williams do grupo de *PDP* do MIT que 1986, divulgou e popularizou o uso do *backpropagation* dando, assim um novo impulso ao desenvolvimento das redes neuronais. Este algoritimo é conhecido como retropropagação do erro, ou ainda, por regra delta generalizada sendo o seu objectivo a minimização do erro quadrático médio.

Com esta descoberta, as **MLP** tornaram-se capazes de solucionar problemas que não são linearmente separáveis. Trata-se de um método muito simples, mesmo para modelos complexos contendo milhares de parâmetros (pesos); as **MLP,** são assim, uma técnica flexível para aferirem padrões estatísticos de reconhecimento com modelos complexos”.

Conforme o seu nome indica, **MLP** são compostas por: Uma camada de entrada (E); uma ou várias camadas escondidas ou intermédios (I); uma camada de saida (S); Um conjunto de conexões unidireccionais (C), definidos pelo iniciais (i, j, w) ou abreviadamente w_{ij}, em que $i \in I \cup S$, $j \in E \cup I$, $j < i$ e $w \in \Re$; e um conjunto de funções de activação (F), normalmente do tipo não linear e diferenciáveis, sendo a função logística (ou sigmoid) uma das mais utilizadas.

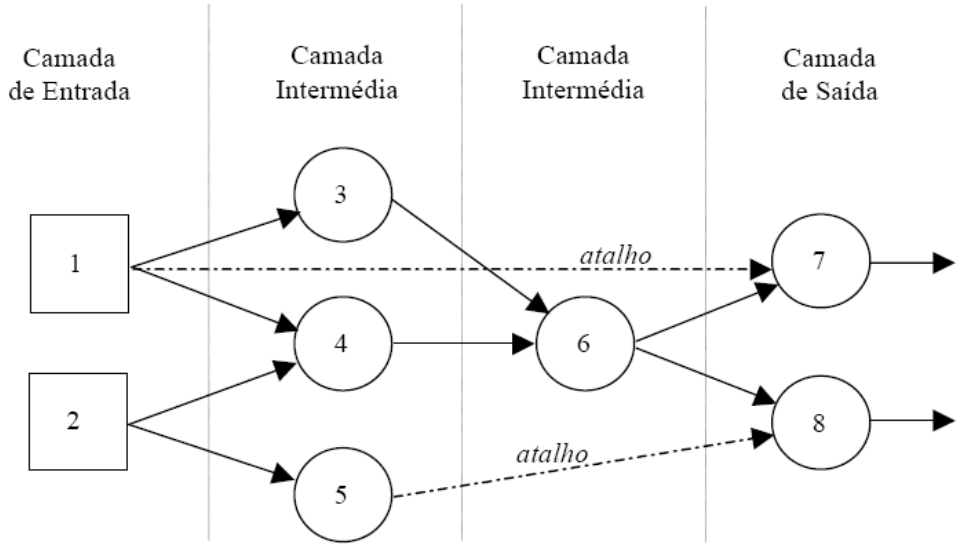


Figura 5.7- Estrutura de um MLP

Os *inputs* são apresentados simultaneamente à camada de entrada, sendo que os *inputs* ponderados desta camada servem de alimentação à camada seguinte (camada escondida) e assim sucessivamente.

Cada uma das camadas possui, uma função especifica. “A camada de saida recebe os estimulos da camada intermédia e constroi o padrão que será a resposta. A camada intermédia funciona como extractoras de caracteristícas, cujos pesos são uma codificação de caracteristicas apresentadas nos padrões de entrada, permitindo que a rede crie a sua própria representação, mais rica e complexa, do problema. Assim, além de calcular o sinal de saida de unidade calcula uma estimativa instantânia do vector gradiente que é necessária para a retropropagação do erro. A camada de entrada é responsavel por receber os dados externos e converter em representação intangível para a rede” (Pag.78).

5.6.3 Algoritimo Backpropagation

Dentro dos algoritmos supervisionados, o algoritmo de *Backpropagation* (BP) é talvez o método de aprendizagem mais popular e mais utilizado em RNAs Este algoritimo de aprendizagem representa um marco na evolução das redes neuronais artificiais, pois que, enunciando (Chorão 2005) “foi o desenvolvimento de um método de retropropagação

do erro que ressuscitou o interesse pelas redes neuronais".

Conforme (Beale 1990), o *backpropagation* pode ser visto como uma generalização do método Delta para redes neurais de múltiplas camadas, sendo que a a principal modificação reside no processo de cálculo e actualização dos pesos durante a fase de treino. A grande dificuldade dos perceptrões de multicamada consiste no cálculo dos pesos nas camadas intermédias duma forma eficiente e que minimize o erro na saída. Quantas mais camadas intermédias tiver, mais difícil se torna o cálculo dos erros. O valor do erro na saída é fácil de calcular, pois é a diferença entre a saída obtida e a saída desejada, mas nas camadas intermédias a dificuldade é acrescida, pois não existe uma observação directa do erro entre as camadas. O algoritmo de Retropropagação veio preencher esta lacuna.

Trata-se de um algoritimo em que a aprendizagem dá-se através de um processamento interactivo dos exemplos de treino, comparando as previsões da rede para cada um dos exemplos de treino com os verdadeiros valores. A minimização do erro no algoritmo *backpropagation* é obtida pela execução do gradiente decrescente na superfície de erros do espaço de pesos, onde a altura para qualquer ponto no espaço de pesos corresponde à medida do erro. Para cada exemplo de treino, os pesos são modificados de forma a minimizar o erro quadrático médio entre as previsões da rede e os verdadeiros resultados. Estas modificações são feitas no sentido contrário, da camada de *output* para a camada de *input*. O erro é apurado na camada de *output* e "retro-propagado" para a camada de *input*, ou seja, uma vez apurado o erro segue-se um processo de "apuramento das responsabilidades" tentando corrigir os pesos que mais contribuíram para esse erro.

Resumindo é possivel identificar duas fases distintas no processo de aprendizagem do algoritimo de retropropogação:

- A primeira fase é responsável pelo processo de treino, e consiste em enviar um sinal funcional que vai da camada de *input* até a de *output*, i.e., processamento para frente, onde um vector de entrada (X^p) é fornecido aos neurónios de entrada, propagando-se para frente, camada a camada. Finalmente, é produzido um conjunto de saída como resposta da rede. Durante a fase de propagação os

pesos sinápticos da rede são todos fixos.

- Na segunda fase do treino é enviado um sinal do erro, no sentido inverso, isto é, de *output* para a camada de *input*- demoninado de retropropagação. Durante a fase de retropropagação, os pesos sinápticos são todos ajustados de acordo com uma regra de correcção do erro. Especificamente esta fase representa a validação da fase anterior, ou seja, verifica-se se o *output* produzido foi satisfatório, através da comparação de saídas geradas pela rede com a resposa desejada para produzir um sinal de erro. Este sinal de erro é também retropropagado através da rede, em sentido contrário das conexões sinápticas – daí o nome retropropagação do erro.

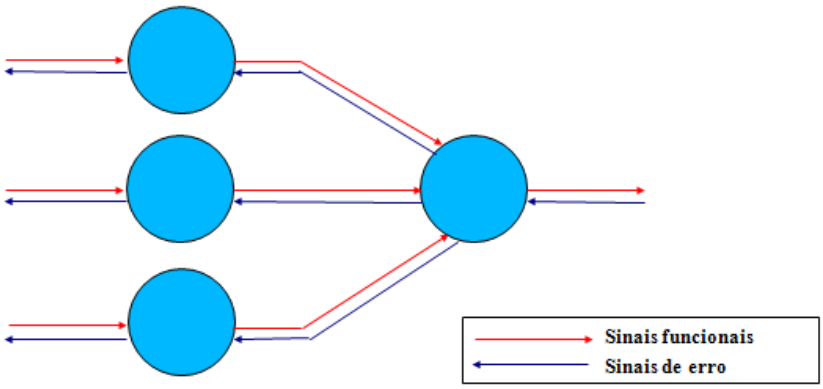


Figura 5.8: Vagas de computação

De modo a facilitar a compreensão do algoritimo, apresenta-se de seguida uma descrição resumida dos passos mais importantes do algoritimo de retropropagação. Para mais detalhes aconselhamos a consulta de (Freeman e Skapura 1992) e (Haykin 1999). Para tal, considere a seguinte arquitetura multicamada apresentada na figura 5.9.

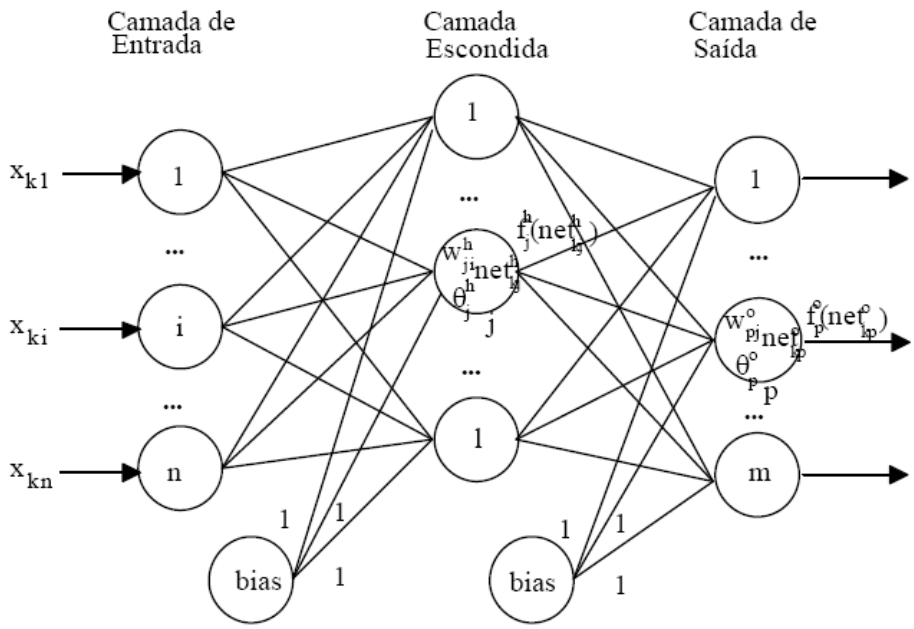


Figura 5.9 Arquitetura de rede mult-icamada

Fonte: (Roisenberg e Vieira, Redes Neuronais Artificiais: Um Breve Tutorial s.d.)

Considere um conjunto de P pares de vetores (X_1, Y_1), (X_2, Y_2),..., (X_P, Y_P), no nosso conjunto de treino e que são exemplos de um mapeamento funcional definido como: $Y = \theta(X) : X \in \Re^n, Y \in \Re^m$. Com o objectivo de treinar a rede de modo que ela consiga aprender uma aproximação da forma: $O = Y' = \theta(X) : X \in \Re^n, Y' \in \Re^m$ devemos seguir os seguintes:

O mapeamento funcional que foi proposto acima pode ser entendido como um conjunto de coordenadas cartesianas em que para cada x_i de entrada existe um y_i de saída. Assim, escolhendo para uma função qualquer um bom exemplo de treino P (x_i, y_i) a rede será capaz, depois de treinada, de interpolar/generalizar novos exemplos, criando assim uma aproximação da função.

Portanto, comforme descrito anteriomente, em primeiro lugar, um vector de entrada X_k= [x_{k1} x_{k2}... x_{kn}] ^{T} do conjunto de treino é apresentado à camada de entrada da rede. Os elementos de entrada distribuem os valores para os elementos da camada escondida. Para calcular o valor do *net* para o j-ésimo elemento da camada escondida, procedemos à

multiplicação dos *outputs* de cada camada anterior pelo respectivo peso (w) e posteriormente a soma de todos eles. A expressão matemática é dada por:

$$net_{kj}^{h} = \sum_{i=1}^{n} w_{ji}^{h} x_{ki} + \theta_{j}^{h} \qquad (5.1)$$

Onde w_{ji} é peso da conexão entre o j-ésimo elemento da camada de entrada e o j-ésimo elemento da camada escondida *h* e θ_{p}^{o} é o termo opcional chamada *bias* que prevê um factor ficticio de entrada igual a 1, dando um grau de liberdade maior para a função de saída do neurónio.

Assumindo que os neurónios são estáticos, assumimos que o valor da função de activação será igual ao *net*, então, o valor de saída para um neurónio da cada escondida resulta da expressão:

$$i_{kj} = f_{j}^{h}(net_{kj}^{h}) \qquad (5,2)$$

Do mesmo modo, as equações para os neurónios da camada de saída são dadas por:

$$net_{kp}^{o} = \sum_{j=1}^{i} w_{pi}^{o} i_{kj} + \theta_{p}^{o} \quad \text{e} \quad o_{kp} = f_{p}^{o}(net_{kp}^{o}) \qquad (5,3)$$

Conforme (Freeman e Skapura 1992), o objectivo do treino consiste em ensinar à rede o mapeamento de todo vector de entrada para o respectivo vector de saída, isto é, encontrar os valores apropriados para os pesos das conexões da rede de modo a minimizar a função do erro definida pela soma dos erros quadráticos médios da rede.

Assim, o erro para um único neurónio *p* na camada de saída para um vector de entrada k é dado por

$$E_{kp} = (y_{kp} - O_{kp}) \qquad (5.4)$$

De forma a minimizar a função de custo, calcula-se a derivada em ordem ao peso sináptico, ou seja, a direcção de modificações dos pesos será dada de acordo com a direcção que o vector gradiente seguir na superfície. Aplicando a regra de cadeia tem-se:

$$\frac{\partial E_k}{\partial w_{pj}^o} = -(y_{kp} - o_{kp}) \frac{\partial f_p^o}{\partial (net_{kp}^o)} \frac{\partial (net_{kp}^o)}{\partial w_{pj}^o} \qquad (5.5)$$

Podemos escrever a derivada de $\partial f_p^o net_{kp}^o$ e o último termo da equação como:

$$\frac{\partial (net_{kp}^o)}{\partial w_{pj}^o} = \frac{\partial}{\partial w_{pj}^o} \sum_{j=1}^{1} w_{pj}^o i_{kj} + \theta_p^o) = i_{kj} \qquad (5.6)$$

Combinando as equações, o negativo do gradiente será:

$$\frac{\partial (E_k)}{\partial w_{pj}^o} = (y_{kp} - o_{kp}) f_p^{o'}(net_{kp}^o) i_{kj} \qquad (5.7)$$

Por aplicação do método de gradiente descendente, poder-se á evidenciar que a alteração do peso sináptico deve dar-se na direcção oposta da derivada da superficie do erro aplicando-se taxa de aprendizagem η , pelo que a alteração deve repeitar:

$$w_{pj}^o(t+1) = w_{pj}^o(t) + \Delta_k w_{pj}^o(t) \text{ Com } \Delta_k w_{pj}^o = \eta (y_{kp} - o_{kp}) f_p^{o'}(net_{kp}^o) i_{kj} \qquad (5.8)$$

Convém ressaltar que a função f_p^o precisa ser uma função diferenciável para que seja possível implementar a busca do gradiente descendente. A funcão logistica ou sigmoidal, pela facilidade de cálculo da sua derivada preencher os requesitos de continuidade, diferenciabilidade e monotonicidade, é a mais utilizada sendo a sua expressão a seguinte:

$$f_p^o(net_{jp}^o) = \frac{1}{1 + e^{-net_{jp}^o}} \text{ E a sua derivada dado por: } f_p^{o'} = f_p^o(1 - f_p^o) \qquad (5.9)$$

Os cálculos para os neurónios das camadas escondidas são similares, salvo facto de não sabermos a *priori* qual o valor desejado de saída para os neurónios destas camadas. Assim, o cálculo é feito em função das saídas desejadas pela camada de saída, pois

estas estão intimamente ligadas com as saídas dos neurónios das camadas intermedédia. Daí vem que;

$$\begin{aligned} E_k &= \frac{1}{2}\sum_p (y_{kp} - O_{kp})^2 \\ &= \frac{1}{2}\sum_p (y_{kp} - f_p^o(net))^2 \\ &= \frac{1}{2}\sum_p (y_{kp} - f_p^o(\sum_j w_{pj}^o i_{kj} + \theta_p^o))^2 \end{aligned} \tag{5.10}$$

Sabendo que i_{pj} depende dos pesos da camada escondida, podemos utilizar este facto para calcular o gradiente de E_k em relação aos pesos da camada escondida.

$$\begin{aligned} \frac{\partial E_k}{\partial w_{ji}^h} &= \frac{1}{2}\sum_p \frac{\partial}{\partial w_{ji}^h}(y_{kp} - O_{kp})^2 \\ &= \sum_p (y_{kp} - O_{kp}) \frac{\partial O_{kp}}{\partial (net_{kp}^o)} \frac{\partial (net_{kp}^o)}{\partial i_{kj}} \frac{\partial i_{kj}}{\partial (net_{kj}^h)} \frac{\partial (net_{kj}^h)}{\partial w_{ji}^h} \end{aligned} \tag{5.11}$$

Cada um dos factores da equação pode ser calculado explicitamente das equações anteriores, assim como foi feito para o gradiente da camada de saída. O resultado fica:

$$\frac{\partial E_k}{\partial w_{ji}^h} = -\sum_p (y_{kp} - O_{kp}) f_p^{o'}(net_{kp}^o) w_{pj}^o f_j^{h'}(net_{kj}^h) x_{ki} \tag{5.12}$$

Por fim, assim como no caso da camada de saída, actualizamos os pesos da camada escondida proporcionalmente ao valor negativo da equação.

$$w_{ji}^h(t+1) = w_{ji}^h(t) + \Delta_k w_{ji}^h(t)$$

Onde

$$\Delta_k w_{ji}^h = \eta f_j^{h'}(net_{kj}^h) x_{ki} \sum_p (y_{kp} - o_{kp}) f_p^{o'}(net_{kp}^o) w_{pj}^o \tag{5.13}$$

Antes de se iniciar o treino de uma rede, há que proceder à escolha dos valores iniciais dos pesos associados às ligações entre nodos, que em geral pertencem ao intervalo [0,1] ou [-1,1] e são gerados de forma aleatória. Os exemplos de treino são apresentados sucessivamente às unidades visíveis da rede neural artificial, até que um erro aceitável (previamente fixado) seja alcançado ou enquanto um número determinado de iterações não for satisfeito. O último conjunto de pesos observado entre as conexões das células é então mantido para testar a habilidade da rede em mapear a função de entrada para saída e a consequente validação do modelo de redes neurais artificiais.

5.6.4 Considerações sobre o Algoritmo Backpropagation

O desempenho do algoritmo de aprendizagem *backpropagation* está condicionado à arquitetura da rede neuronal e ao conjunto de dados utilizados no processo de ajustamento dos pesos sinápticos entre as conexões da rede.

5.6.4.1 Dados de treino

Conforme James Freeman (Freeman e Skapura 1992) não existe critério específico para selecção dos exemplos de treino. É possível utilizar todos os dados disponíveis no processo de treino da rede, embora apenas um subconjunto desses dados seja suficiente para que o processo seja executado com sucesso. Os dados restantes podem ser usados para avaliar a capacidade de generalização do *backpropagation*. Idealmente, os dados devem ser em quantidade suficiente para reflectir todas as possíveis variações de respostas diferentes. Neste caso, os dados dividem-se em três conjuntos: um conjunto de treino, que servirá para a actualização dos pesos das sinápses; um conjunto de teste, que serve para verificação da resposta da rede a dados não usados para treino; e um conjunto de validação, que deve ter casos diferentes dos anteriores, e permitirá aferir qual a melhor rede obtida pelo treino.

5.6.4.2 Tipo de processamento, taxa de aprendizagem e mínimos locais

Conforme a descrição anteriormente efectuada, os pesos sinápticos vão sendo ajustados no decorrer do processo de treino. No algoritimo *backpropagation*, este ajustamento pode

ser efectuado através de um processamento em modo sequencial (por padrão ou ainda em *on-line*) ou em *batch,* também denominado de processamento por ciclo.

No processamento em modo sequencial os pesos da rede são actualizados à medida que um novo exemplo de treino {entrada, saída} é apresentado à rede. O treino sequencial é muito utilizado em aplicações em tempo real, devido ao facto de utilizar menos memória no seu processamento uma vez que os padrões são apresentados à rede par a par e os pesos são actualizados após o seu processamento. Este facto faz com que a rede tenha uma maior probabilidade de não cair num mínimo local, bem como, seja mais difícil estabelecer condições teóricas para a convergência do algoritmo. Uma das vantagens da utilização do método sequencial consiste no facto de, ao trabalhar com dados extensos e redundantes, o algoritmo conseguir tirar partido, já que os dados são apresentados à rede par a par. Apesar das desvantagens do modo sequencial em detrimento do modo em *batch*, o mesmo é muito usado devido ao facto de ser um algoritmo simples de aplicar e proporcionar soluções em vários tipos de problemas com dificuldades diversas (Haykin 1999)

No processamento em *batch*, a actualização dos pesos é realizada após todos os exemplos de treino {entrada, saída} serem apresentados à rede e processados em conjunto formando uma época. O treino em *batch* melhora a estimativa do vector de gradiente sendo a convergência para um mínimo local garantida através do uso de condições simples. Este modo de processamento permite mais facilmente estabelecer comparações entre os diversos parâmetros escolhidos.

O algoritimo de retropropagação apresenta, contudo, alguns problemas tais como a paralesia da rede e a existência de um minimo local. Como se sabe, o algoritimo de retropropagação utiliza a heurística do gradiente decrescente para ajustar os pesos entre as sinápses, seguindo a curva da superfície dos erros em direcção a um ponto mínimo (Wasserman 1989).As superfícies de erros convexas, por apresentarem um único mínimo, permitem que este método atinja o mínimo global. Nas superfícies de erros não convexas e altamente convolutas, normalmente encontradas em problemas práticos, a solução alcançada pode não ser a óptima. Nestes casos, haverá que ser utilizado algum algoritmo de optimização global.

Assim que um mínimo é encontrado, seja global ou local, a aprendizagem termina (Freeman e Skapura 1992). Se a rede alcançar um mínimo local (figura 5.10), todas as direcções na sua vizinhança mais próxima representam valores maiores que o alcançado e, consequentemente, a convergência para o mínimo global não é atingido. Nesse caso, a magnitude do erro da rede pode ser muito alta e, portanto, inaceitável.

Caso a rede neural encerre a aprendizagem antes que uma solução satisfatória seja obtida, o redimensionamento do número de unidades ocultas ou da taxa de aprendizagem e do termo *momentum* podem ser suficientes para resolver o problema, como se explicará mais adiante.

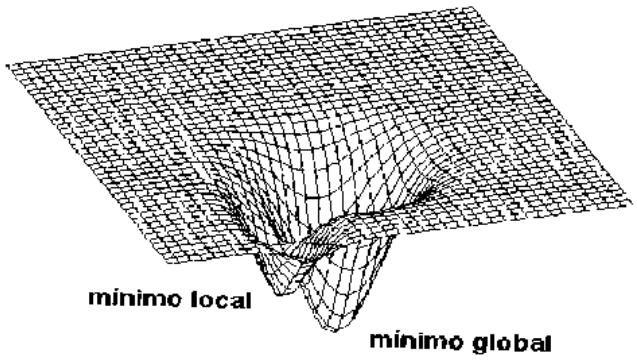


Figura -5.10 Minimo Local

Fonte : (Kovacs 1996)

Através da figura 5.10, que ilustra um corte em uma superfície de erros hipotética no espaço de pesos, é possível observar um ponto de mínimo local. Tanto à direita, quanto à esquerda, os valores são maiores que esse mínimo.

(Freeman e Skapura 1992), sugere que os pesos das conexões entre as camadas de uma rede neural sejam inicializados com valores aleatórios e pequenos para que se evite a saturação da função de activação e a consequente incapacidade de realizar a aprendizagem. Quanto mais pequena for a taxa de aprendizagem η, menores vão ser as mudanças nos pesos das conexões da RNA, de modo que a procura do mínimo global será favorecida pelo uso de saltos mais suaves. O problema que se coloca é que, desta forma converge-se para uma aprendizagem mais lenta. Por outro lado, se se aumentar em demasia o valor de η, então os saltos desmedidos nas mudanças dos pesos poderão

provocar instabilidade no treino (e.g., movimento oscilatório). À medida que o treino evolui, os pesos sinápticos podem passar a assumir valores maiores, forçando a operação dos neurónios na região onde a derivada da função de activação é muito pequena. Como o erro retropropagado é proporcional a esta derivada, o processo de treinamento tende a se estabilizar, levando a uma paralisação da rede sem que a solução tenha sido encontrada. Isto pode ser evitado pela aplicação de uma taxa de aprendizagem menor. Teoricamente, o algoritmo de aprendizagem exige que a mudança nos pesos seja infinitesimal. Entretanto, a alteração dos pesos nessa proporção é impraticável, pois implicaria um tempo de treino infinito. Por este facto, é recomendável que a taxa de aprendizagem assuma um valor maior no início do treino e, à medida que se observe decréscimo no erro da rede, essa taxa também seja diminuída. Como refere (Beale 1990), à medida que a taxa de actualização dos pesos diminui, o gradiente decrescente torna-se mais apto a alcançar uma solução melhor. Uma forma de aumentar a velocidade de convergência da rede neuronal artificial é a adopção de um método chamado *momentum*. O propósito deste método consiste em adicionar, aquando do cálculo do valor da mudança do peso sináptico, uma fração proporcional à alteração anterior. A equação (5.15) especifica o ajustamento das conexões entre unidades de processamento pela aplicação do termo *momentum.* Outra forma distinta para lidar com este problema reside no uso de diferentes taxas de aprendizagem, uma por cada nodo. (Yeung 1999) sugere que se utilize. $\eta = \frac{1}{\sqrt{z}}$, Para um nodo com z conexões

$$. \; w_{pk}^{0}(t-1) = w_{pj}^{0}(t) + \alpha \Delta w(t-1) \qquad (5.14)$$

Onde α representa o termo *momentum*, $0 < \alpha < 1$.

Na Fig. 5.11 pode-se analisar o comportamento do algoritmo sem e com o termo momento, donde facilmente se percebe a razão pela qual o termo momento ajuda no processo de actualização dos pesos.

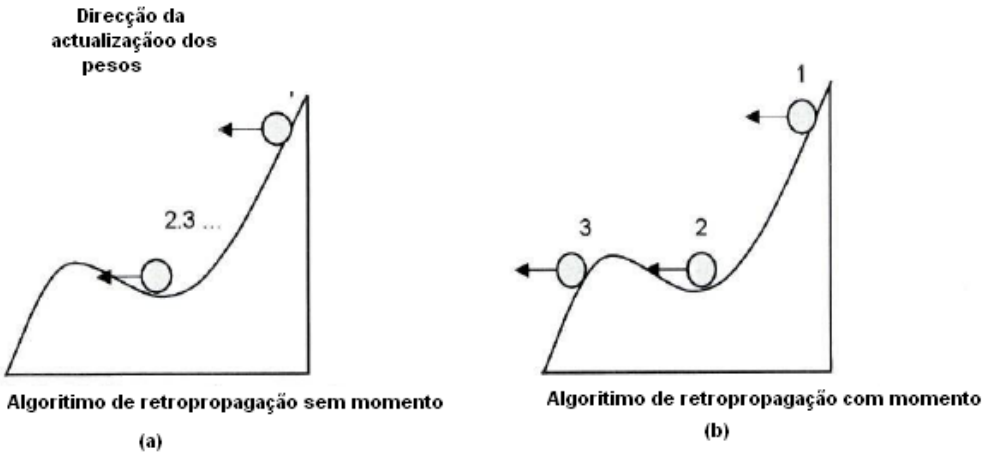


Figura 5.11 – Comportamento do algoritimo de retropropagação com sem e com o termo momento

Fonte: (Meneses 2003)

A introdução do termo momento no algoritmo de Retropropagação representa uma pequena modificação na actualização dos pesos. O termo momento tende a acelerar a convergência da rede evitando a oscilação da mesma e a sua "queda" num mínimo local da superfície de erro. O algoritmo de Retropropagação assume que a taxa de aprendizagem é constante, mas, usando o termo momento, tal parâmetro é variável (depende da conexão da rede).

5.6.4.3 Número de camadas. Número de neurónios

Um dos problemas enfrentados no treino de redes neuronais do tipo MLP diz respeito à definição do número de camadas e neurónios. A definição destes parâmetros é um processo tão pouco compreendido que são muitas vezes chamadas de "magia negra". Pequenas diferenças nestes parâmetros podem levar a grandes diferenças tanto no tempo de treino como na generalização obtida.

Exitem inúmeras pesquisas no sentido de encontrar uma fórmula "mágica" que determine a configuração ideal de uma rede neural para um dado problema. No entanto, até agora o que se tem são apenas sugestões que estão fundamentadas em experiências vividas por especialistas no assunto e no bom senso.

Deve-se ter em mente que é preciso obter um modelo que não seja muito rígido a ponto de não modelar fielmente os dados, mas que também não seja excessivamente flexível a ponto de modelar também o ruído presente nos dados. A idéia é que a rede responda de acordo com as características presentes nos dados de entrada e não exatamente igual aos dados de entrada. Por exemplo, o princípio de Ockham, diz que “deveremos preferir modelos simples a complexos e esta preferência deverá aplicar-se até que os modelos se adequem aos dados”. Igualmente, (Chorão 2005) refere que “apesar de várias prácticas para determinar a dimensão da camada inetermédia, na maioria dos casos continua a ser a tentativa e erro a melhor regra a seguir”.

De acordo, (Bação 2005) uma rede MLP com uma camada intermédia é suficiente para aproximar qualquer função contínua e em problemas excepcionalmente complexos se podem utilizar duas. Independentemente da complexidade do problema, duas camadas são suficientes para que a rede possa aproximar o problema.

A utilização de um grande número de camadas escondidas não é recomendada. Cada vez que o erro médio durante o treinamento é utilizado para actualizar os pesos das sinápses da camada imediatamente anterior, ele se torna menos útil ou preciso. A única camada que tem uma noção precisa do erro cometido pela rede é a camada de saída. A última camada escondida recebe uma estimativa sobre o erro. A penúltima camada escondida recebe uma estimativa da estimativa, e assim por diante.

Em relação ao número de neurónios nas camadas escondidas, este é geralmente definido empiricamente. Deve-se ter cuidado para não utilizar nem unidades demais, o que pode levar a rede a memorizar os dados de treino (*overfitting*), ao invés de extrair as características gerais que permitirão a generalização, nem um número muito pequeno, que pode forçar a rede a gastar tempo em excesso tentando encontrar uma representação óptima. Devido a estas dificuldades é recomendado manter o número de neurónios escondidos baixo, mas não tão baixo quanto o estritamente necessário.

Existem várias propostas de como determinar a quantidade adequada de neurónios nas camadas escondidas de uma rede neural. As mais utilizadas são:

- O número de neurónios deverá estar compreendido entre o número de variáveis de *input* e o número de *output* (Blum 1992, 60)
- O numero de neurónios deverá ser menor que a metade do número de variáveis da primeira camada (Swingler 1996, 53).
- O número de neurónios deverá ser igual ao número de dimensões (componentes principais) necessárias para explicar 70 a 90% da variabilidade dos dados de entrada. (Boger e Guterman 1997)
- Utilizar um número de sinápses dez vezes menor que o número de exemplos de treino disponíveis. Se o número de exemplos for muito maior que o número de sinápses, *overfitting* é improvável, no entanto pode ocorrer *underfitting* (a rede não converge durante o processo de treino).

5.6.4.4 Generalização overfitting

Um aspecto tido como fulcral aquando da elaboração de um modelo para *credit scoring*, seja o modelo neuronal seja o modelo logit, (ou no desenvolvimento de parte dos métodos preditivos não-paramétricos) prende se com a sua capacidade de generalização, isto é, qual a qualidade das previsões produzidas pela rede para casos que não se encontrem no conjunto de dados de treino?

Diz-se que uma RNA possui uma boa generalização quando a correspondência entre entradas e saídas é correcta (ou próxima disso) para dados de teste, retirados da mesma população, nunca antes utilizados na criação ou treino da rede. O processo de aprendizagem pode ser visto como um problema de ajustamento de curvas ou de aproximação de funções, onde a rede tenta efectuar uma boa interpolação não linear dos dados (Riedmiller e Braun. 1993). A Figura 5.12 mostra como podem ocorrer duas generalizações distintas para o mesmo conjunto de dados de treino. Aqui, uma boa generalização ocorre com a curva A, com um erro mínimo para os dados de teste. O mesmo já não sucede com a curva B, que origina um erro maior para os casos de teste, isto apesar de apresentar um menor erro para os dados de treino. Tal fenómeno, designado de *overfitting,* ocorre quando uma RNA memoriza em demasia os exemplos de treino, tratando-se de um dos problemas mais sérios relacionados com o uso de RNAs (Russel e Norvig 1995). Durante o processo de aprendizagem, a rede pode captar certas

características, como o ruído, que estão presentes nos dados de treino, mas não na função implícita a ser aprendida. Este exemplo ilustra os dois objectivos contraditórios da aproximação funcional. Por um lado tem-se a minimização do erro de treino, pelo outro tem-se a minimização do erro para as entradas desconhecidas. Assim, uma RNA que seja treinada em demasia perde capacidade para generalizar.

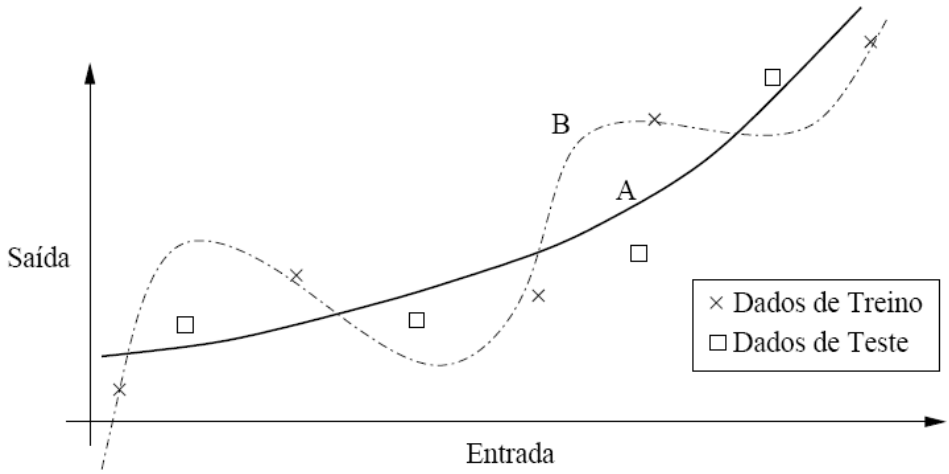


Figura 5.12- Generalização e Overfitting

A generalização nem sempre é possível. Existem 3 condições que são normalmente necessárias (nem sempre suficientes) para uma boa generalização:
(Gallant 1993) e (Sarle, Stopped Training and Other Remedies for Overfitting. 1995):

- A primeira condição está associada à complexidade do problema a ser aprendido – Trata-se de um factor de difícil controlo. As entradas devem conter informação suficiente para permitir a obtenção das saídas desejadas; i.e, tem de existir uma função matemática com algum grau de precisão que relacione as entradas com as saídas. Por outro lado, convém que esta função seja suave; i.e, pequenas alterações nas entradas devem provocar pequenas alterações nas saídas, para a maior parte dos casos. Por vezes, uma transformação não linear nas entradas pode melhorar a sua suavidade (transformação logarítmica);

- A segunda condição para uma boa generalização é a de que o conjunto de exemplos de treino seja suficientemente grande e representativo da população A generalização é sempre efectuada a partir de dois tipos de duas situações:

interpolação e extrapolação. No primeiro caso, um valor é calculado a partir da informação dos valores constantes de casos na vizinhança. A segunda situação engloba tudo o resto, ou seja, casos fora do domínio dos dados de treino. Enquanto a interpolação pode ser efectuada com relativa acuidade, o mesmo já não se passa com a extrapolação, notoriamente menos fiável;

- A terceira condição tem que ver com a arquitectura da RNA; i.e, o número de parâmetros livres que denotam os pesos das ligações entre neurónios e a sua capacidade de aprendizagem bem como a sua complexidade. Uma rede não propriamente complexa irá falhar na aproximação à função a aprender. Por outro lado, uma rede demasiado complexa, irá fixar o ruído existente nos dados, provocando *overfitting*. A Figura 5.13 mostra uma variação típica do erro de uma RNA com uma camada intermédia, para os casos de treino e de teste, com o incremento do número de neurónios intermédios. À medida que estes aumentam o erro de treino diminui. A dada altura, a curva de erro para os casos de teste inflecte, perdendo-se em generalização.

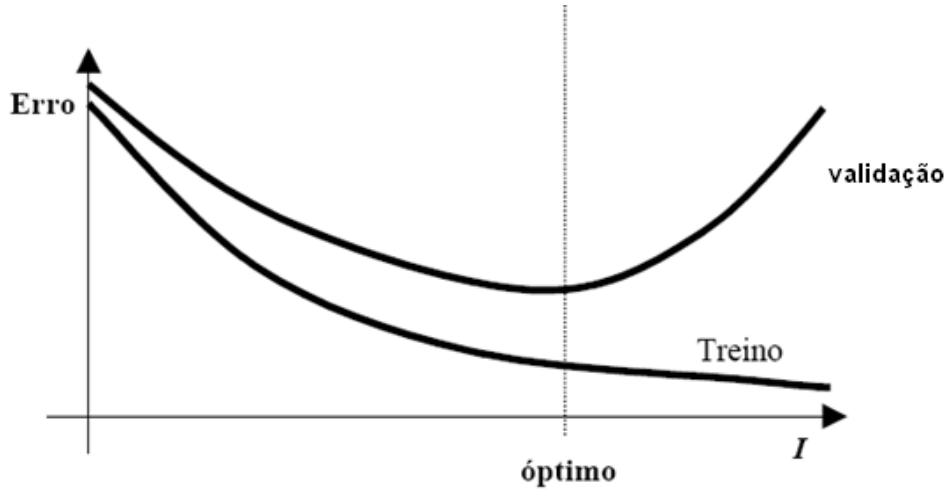


Figura 5.13- Erro típico que ocorre com o aumento do número de neurónios da camada intermédia

A melhor forma de evitar o *overfitting* é utilizar uma quantidade elevada de exemplos de treino. Quando este número for pelo menos 30 vezes superior ao número de conexões, então é pouco provável que ocorra *overfitting*. O problema que surge é que nem sempre existem muitos exemplos de treino disponíveis e não se deve reduzir o número de

conexões de um modo arbitrário, devido a problemas de insuficiência de complexidade da rede. Dada uma dimensão fixa de dados de treino, existem pelo menos duas grandes alternativas eficientes para evitar a sub-aprendizagem e a sobre-aprendizagem, permitindo assim uma boa generalização (Sarle, Stopped Training and Other Remedies for Overfitting. 1995), (Sarle, Neural network 1999): Regularização e selecção de modelos.

5.6.4.4.1 Regularização

A regularização baseia-se num controlo dos valores dos pesos das conexões da rede para se obter uma boa generalização. Entre os diversos métodos de regularização apresentamos os definidos por (Russel e Norvig 1995) (Sarle, Stopped Training and Other Remedies for Overfitting. 1995):

Decaimento de pesos:

A estratégia passa por acrescentar uma penalidade à função de erro, de modo a reduzir os pesos das conexões, em particular as mais expressivas, visto que estas prejudicam o processo de generalização, dando origem a funções irregulares, por vezes na vizinhança de descontinuidades. Por outras palavras, pesos cujo $|\bullet| >> 0$ causam uma excessiva variância nas saídas, (onde $|\bullet|$ denota a função módulo). (Barlett 1997). Normalmente esta penalidade é dada pela expressão: $d \times \sum w_{ij}^2$ onde d representa a constante de decaimento, cuja escolha é crucial para uma boa generalização.

Adição de ruído

O objectivo é acrescentar deliberadamente ruído artificial às entradas durante o treino. Esta estratégia funciona porque a maior parte das funções a serem aprendidas pela rede são suaves. Assim, em cada iteração do algoritmo de treino, novos casos de treino são criados, acrescentando ruído. Este nem deve ser demasiado pequeno, produzindo pouco efeito, nem demasiado grande, pois obviamente desvirtuará a função implícita a ser aprendida. Este ruído é produzido por um gerador de números aleatórios, usualmente seguindo uma distribuição normal com média 0 e desvio padrão s, cujo valor deverá ser

estimado de algum modo (e.g., de modo a que seja menor do que o erro de generalização, medido por um estimador).

Paragem antecipada

Trata-se de um dos mais populares métodos de regularização, onde os dados de treino são divididos em dois tipos de casos: de treino e de validação. Os primeiros são utilizados na aprendizagem da rede, enquanto os últimos são utilizados para aferir a qualidade da aprendizagem; i.e., para estimar o erro de generalização. De notar que podem ser utilizados novos casos de teste para medir o desempenho da rede após o treino. Durante a fase de treino, calcula-se o erro de validação de forma periódica, parando-se quando este começa a aumentar. Todavia, esquemas de paragem mais elaborados têm de ser adoptados, dado que a função de erro pode apresentar diversos mínimos locais. Por exemplo, Prechelt (Patterson 1996) defende o uso de três critérios de paragem:

- O primeiro critério é denominado de falha no processo de treino, que consiste em avaliar o progresso do treino, isto é, a diminuição do erro sobre os exemplos de treino, ξ_{tr}, durante uma dada faixa do treino, com k iterações. A função de progresso, avaliada em cada k iterações, toma a forma:

$$p_k(t) = 1000 \times \left(\frac{\sum_{t' \in t-k+1 \ldots t} \xi_{tr}(t')}{k \times \min_{t' \in t-k+1 \ldots t} \xi_{tr}(t')} - 1 \right) \qquad (5.16)$$

O progresso no treino é elevado nas suas fases de maior instabilidade, onde o erro para os exemplos de treino sobe em vez de diminuir. No entanto, tende para zero a longo prazo, a não ser que o treino se torne oscilante. O treino é parado se $p_k(t) < \beta$, em que β é uma medida de erro em estado estacionário;

- Perda de Generalização - Esta ocorre sempre que há uma inversão de sinal nos valores da derivada da função de erro para os casos de validação, ξ_{va} passando estes de negativos a positivos. A função de avaliação, também medida de k em k iterações, toma a forma:

$$G_k(t) = 100 \times \left(\frac{\xi_{va}(t)}{\min_{t' \leq t} \xi_{va}(t')} - 1 \right) \tag{5.17}$$

Uma grande perda de generalização é uma boa razão para se parar o treino, pelo que o treino termina se $G_k(t) > \alpha$, onde α denota a perda de poder de generalização aconselhável para a rede; e

- Número Máximo de Iterações – Este critério é aplicado quando os anteriores critérios de paragem falham, de modo a garantir que o treino termine.

A paragem antecipada é bastante utilizada porque é simples e rápida, podendo ser aplicada a RNAs com um grande número de conexões. Todavia, possui algumas desvantagens. Em primeiro lugar, é bastante sensível à forma como é feita a divisão entre exemplos de treino e de validação; i.e., quais e quantos casos devo usar conjunto. Por outro lado, não aproveita toda a informação disponível para a aprendizagem.

5.6.4.4.2 Selecção de Modelos

A regularização diminui o efeito de *overfitting* pelo estímulo dado à aprendizagem de funções suaves. No entanto, utiliza uma estrutura fixa, que deve ser especificada pelo utilizador. Embora se possa utilizar uma grande estrutura, com um grande número de neurónios intermédios, na prática, a optimização dos pesos torna-se de difícil ajustamento, exigindo um grande esforço computacional. Mais ainda, em geral, são métodos que exigem um delicado balanço, controlado por um (ou mais) parâmetro (s) de regularização. Mais recentemente, métodos Bayesianos têm sido incorporados na regularização, para eliminar alguns destes problemas. Trata-se de uma abordagem promissora embora ainda pouco desenvolvida. Para, além disso, estes métodos assumem certos tipos de distribuição entre dados de treino e teste que podem falhar quando o número de conexões da rede é grande, quando comparado com a cardinalidade dos dados de treino (Kosko 1988). Uma alternativa distinta baseia-se na procura de uma topologia para uma RNA, em termos do número de conexões, número de nodos e camadas intermédias. Os defensores desta estratégia argumentam que é mais fácil

adaptar a complexidade da rede ao problema a ser resolvido. Assim, um problema que seja de difícil aprendizagem para uma rede poderá ser facilmente aprendido por outra rede.

A abordagem estatística à resolução deste problema passa pela estimativa do erro de generalização para cada um dos modelos, ou topologias de rede, escolhendo-se o modelo que minimiza essa estimativa.

Existem diversos métodos para estimar a capacidade de generalização de uma RNA (Efron e Tibshirani 1993) (Kernsley e Martinez s.d.) alguns dos quais são enunciados a seguir:

- Estatísticas Simples – Diversas métricas foram desenvolvidas tendo em conta modelos lineares, baseando-se em suposições sobre as amostras

Entre estas, podem-se referenciar:

  - *Critério de Informação de Akaike* - conhecido por AIC. A formulação matemática é dado por

$$AIC = n\ln(SSE/n) + 2p \tag{5.18}$$

Onde SSE representa o somatório do quadrado dos erros para todos os casos de treino, n representa o número de exemplos de treino e p o número de parâmetros livres da rede, ou seja, o número de pesos das ligações entre os neurónios da rede; e

  - *Critério de Informação de Bayes*, designado por BIC ou SBC, que normalmente funciona bem com RNAs.

$$BIC = n\ln(SSE/n) + p\ln(n) + p \tag{5.19}$$

- Validação com Divisão da Amostra – O método mais popular para a estimação do erro de generalização de uma RNA, geralmente associado à paragem antecipada

do treino da rede, baseia-se numa divisão dos dados do problema em casos de treino, para a rede aprender, e casos de validação, para estimar o erro de validação. Como ponto forte deste processo tem-se a sua simplicidade e rapidez, embora produza uma redução efectiva dos casos disponíveis para treino.

5.7 Redes Neuronais e modelos econométricos

Ao contrário do que pode parecer à primeira vista, os modelos de redes neuronais têm vários pontos de contacto com os modelos econométricos tradicionais, nomeadamente os modelos de regressão. Muitas das semelhanças existentes ficam embotadas pelo uso de jorgões técnios diferentes pelos estaticistas ou econometristas e conexionistas.

No caso de modelos de regressão, por exemplo, temos uma variável dita endógena sendo explicada por diferentes variáveis exógenas. Nas redes neuronais as variáveis exógenas podem ser vistas como os neurónios da camada de entrada, enquanto a variável endógena é representada pelo sinal de saida desejável pela rede. Em outras palavras a variável endógena é o padrão que é objecto de aprendizagem da rede neuronal. Na verdade, uma rede neuronal artificial constituida por apenas uma camada de entrada e outra de saída (*perceptron*) pode ser facilmente relacionada com o modelo de regressão linear.

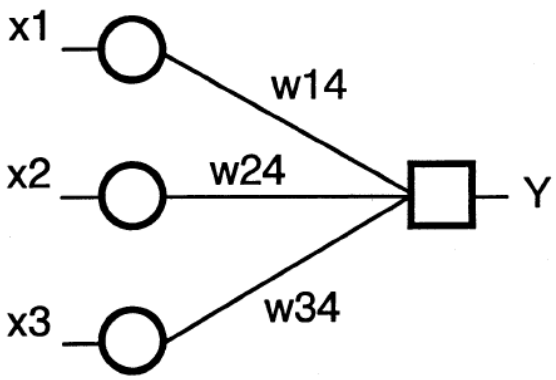


Figura -5.16 Rede Neuronal Artificial

Neste caso, o vector de pesos da rede neuronal da figura 5.16 (w_{14}, w_{24}, w_{34}), sem camada oculta, nada mais é que um vector de parâmetros da regressão. Eles indicam, assim como os parâmetros da regressão, a importância de cada sinal de entrada na explicação do padrão de saída. Contudo, quando utilizamos uma camada oculta, (*perceptron* multi-camada), como é conhecido na literatura de redes neuronais artificiais, estamos a introduzir não-linearidade nas relações entre as entradas x_1, x_2, x_3, e a saída y. Portanto, a relação entre as variáveis endógenas e exógena deixa de ser linear, e a comparação, então, tem que ser feita com modelos de regressão não lineares.

Nos modelos económetricos tradicionais, os parâmetros do modelo são obtidos através de algum processo de estimação que envolve a minimização da soma do quadrado dos resíduos ou a maximização da função de verossimilhança. Já os pesos da RNA, são obtidos, segundo o jargão das redes neuronais, por um processo de aprendizagem. Contudo, a estimação dos pesos dos RNA, a partir de um processo de aprendizagem, e a estimação dos coeficientes dos modelos tradicionais, tal como o modelo logístico, são, do ponto de vista estatístico e matemático, exactamente a mesma coisa. Em ambos os casos, o que se procura é minimizar a função do erro médio quadrático, ou qualquer outra função objectivo escolhida. No caso das RNAs o *backpropagation* é apenas um algoritimo númerico utilizado para obter um minimo local/global da função objectivo em questão. Neste sentido, o processo de aprendizagem das RNAs, é equivalente à estimação dos parâmetros realizada nos modelos enconométricos.

A principal diferença entre os modelos economètricos tradicionais e as RNAs, tal como as conhecemos actualmente, é que estas não contam com uma base estatística pré-determinada. Enquanto nos modelos económetricos estamos a lidar com variáveis aleatórias que possuem uma determinada distribuição conjunta de probabilidade, nas RNA temos apenas sinais de entrada e saída de rede. A ausência de um modelo estatístico bem especificado impede, por exemplo, a construção de intervalos de confiança para estimativas geradas pelas redes neuronais. As previsões geradas pelas redes neuronais são sempre pontuais, ao contrário do que ocorre com os modelos económetricos.

5.8 Principais vantagens e limitações das Redes Neuronais

A utilização das redes neuronais em *credit scoring* pode justificar- se pelas vantagens que estes modelos trazem relativamente a outros métodos de preditivos (Shachmurove 2002), por exemplo, enumera algumas vantagens face aos modelos econométricos tradicionais. Uma das vantagens referidas é que estes modelos conseguem captar com precisão padrões complexos existentes nos dados. A este propósito, (Massoumi, Khotanzad e Abay 1994) mencionam que os dados utilizados nestes modelos são dinâmicos por natureza, sendo necessária a existência de ferramentas não lineares para captar padrões comportamentais existentes nos dados. Outra vantagem, talvez das mais relevantes, é que não é necessário elaborar hipóteses relativamente à natureza da distribuição dos dados. Em vez disso, estes modelos utilizam os próprios dados para produzir representações internas das relações entre as variáveis. Como consequência, é de esperar a obtenção de melhores resultados com a aplicação das redes neuronais quando as relações entre as variáveis não seguem um comportamento pré-definido.

Relativamente às desvantagens, (Shachmurove 2002) refere que não existe uma metodologia estruturada que permita efectuar a melhor escolha relativamente à arquitectura da rede, ao treino da rede ou à verificação da qualidade da mesma. Por exemplo, o número de camadas a utilizar na rede ou o número de neurónios que cada camada deve ter são normalmente aspectos *inter alia* com opiniões divergentes. (Thawornwong e Enke 2004)) para concluírem que as redes neuronais permitem obter melhores resultados que outros modelos na previsão das rendibilidades de acções, compilaram numa tabela as técnicas de modelização neuronais encontradas em 45 estudos diferentes. Verifica-se que raramente há consenso em relação à escolha dos diversos aspectos, o que permite concluir que a escolha da melhor arquitectura para a rede continua a depender em grande medida da sensibilidade e das experiências realizadas por cada investigador.

Outra crítica que se faz amiúde aos modelos neuronais é a crítica da “caixa negra” (*black box*), segundo a qual não é possível perceber como são estimadas as relações nos neurónios ocultos. (Eisinga, Franses e Dijk 1997) Mostraram que esta crítica é um pouco exagerada ao conseguirem, por um lado, desenhar uma rede que lhes permitiu controlar o

timing de activação dos neurónios ocultos e, por outro lado, efectuar inferência relativamente ao impacto que as variáveis independentes provocam nas dependentes. Não obstante, enquanto nos modelos econométricos lineares a avaliação da relevância das variáveis independentes, e do impacto que estas provocam na dependente, é trivialmente avaliada recorrendo às estatísticas t, nos modelos neuronais essa tarefa afigura-se mais complexa.

6 Resultados da estimação dos modelos

6.1 Regresssão Logística

Foram elaborados três modelos logit de acordo com a percentagem escolhida para o conjunto de treino, validação e teste. Os modelos foram estimados por recurso ao *software* SAS (versão 9.1.2), em especial o módulo *Enterprise Miner* (versão 5.2).

De Seguida, apresentamos os resultados e a panóplia de testes estatísticos associados ao modelo logit para o conjunto de treino seleccionado a 70%. Os resultados para o conjunto (60% e 80%.) estão apresentados nos apêndices (A e A1).

Foi estimado um modelo logit binário com as variáveis descritas na pag. 35. Interpôs-se uma variável frequência a fim de balancear o grupo dos *defaults* com o grupo de regulares.

Como se observa na tabela seguinte, o teste que compara o modelo completo, com os 15 parâmetros, com o modelo somente com a constante é estatisticamente significativo, indicando que os parâmentros, quando tomados em conjunto, diferenciam entre clientes em *defaults* e clientes regulares.

Estatística	Qui-Quadrado	g.l	Sig.
Rácio de Verosimilhança	329	14	<.0001
Score	302	14	<.0001
Wald	259	14	<.0001

Tabela 6.1 - Teste de significância para o modelo geral

Coeficiente de determinação – Pseudo R^2

A variância explicada associada aos *defaults* é considerada satisfatória, conforme mostra a tabela 6.2.

R-Square	0.1920	Max-rescaled R-Square	0,21

Tabela 6.2 Coeficiente de determinação.

Qualidade do ajustamento

A tabela 6.3 apresenta os coeficientes de regressão, as estatísticas de Wald, *odds-ratio* e respectivos intervalos de confiança para cada um dos 15 parâmetros. de acordo com o critério de Wald, todos os paramêntros submetem-se à exigência de um nível de significância de 5%, estimados pelo método de *stepwise*.

Variáveis	DF	$\hat{\beta}$	S.E	Wald	Sig	OR	IC para OR a 95%	
							Limite Inferiror	Limite Superior
Intercept	1	0,005	0,045	14,00	0,001	2,824	1,91	7,98
X1	1	10,383	0.4622	50,46	0,001	2,824	1,142	6,99
X2	1	11,261	0.8665	16,89	0,001	3,083	0.564	16,85
X3	1	-0.8859	0.4864	33,17	0,001	0.412	0.159	1,07
X4	1	0.8881	0.1163	58,28	0,001	2,43	1,935	3,05
X5	1	0.8738	0.0986	78,62	0,001	2,396	1,975	2,91
X6	1	-0.6166	0.8012	59,23	0,001	0.540	0.112	2,60
X7	1	0.6794	0.1532	19,66	0,001	1,973	1,461	2,66
X8	1	0.6871	0.3136	48,01	0,002	1,988	1,075	3,68
X9	1	0.4173	0.4472	87,05	0,003	1,518	0.632	3,65
X10	1	0.4475	0.1849	58,55	0,002	1,564	1,089	2,25
X11	1	0.3880	0.2652	21,41	0,002	1,474	0.877	2,48
X12	1	0.2166	0.2088	10,76	0,002	1,242	0.825	1,87
X13	1	0.1807	0.1647	12,03	0,002	1,198	0.867	1,66
X14	1	-0.1409	0.7656	3,39	0,001	0.869	0.194	3,90

Tabela 6.3 - coeficientes de regressão, as estatísticas de Wald,

O teste de Hosmer – Lemeshow, apresentado nas tabelas seguintes, configura mais uma vez uma boa aderência dos dados à realidade observada.

Grupo	Total	Regulares		*Defaults*	
		Obeservados	Esperados	Obeservados	Esperados
1	232	41	41	191	191
2	232	65	68	167	164
3	232	94	89	138	143
4	232	103	100	129	132
5	232	111	110	121	122
6	232	117	121	115	111
7	232	136	132	96	100
8	232	145	145	87	87
9	232	153	161	79	71
10	235	194	191	41	44

Tabela 6.4 -Partição do teste Hosmer Lemeshow.

Chi-Square	*DF*	*Pr > ChiSq*
27688	8	0.9480

Tabela 6.5 - Hosmer Lemeshoe test.

Residuos

Por fim a distribuição dos resíduos é-se confrontado com a plausibilidade de assumirem uma distribuilção normal, encontrando-se 95% das obsrvações entre -1.96 e +1.96.

Teste	Valor	p-value
Kolmogorov-Smirnov	18,5	<.0001
Cramer-von Mises	27	<.0050
Anderson-Darling	15	<.0050

Tabela 6.6 - análise de redisuos

Dbeta

Ordenou-se de forma decrescente, o ficheiro de dados e, apresentam-se as primeiras 16 observações, constatando-se que nenhuma delas é maior que 1, pelo que, na estimação do modelo, não estão incluídas observações tidas como *outliers*.

Dbeta X1	Dbeta X2	Dbeta X3	Dbeta X4	Dbeta X5	Dbeta X6	Dbeta X7	X8	Dbeta X9	Dbeta X10	Dbeta X11	Dbeta X12	Dbeta X13
-0.051	0.043	-0.001	-0.002	-0.258	-0.010	0.013	-0.002	0.025	-0.008	0.001	-0.018	0.034
0.012	0.037	-0.021	-0.002	-0.002	0.004	0.002	0.001	0.006	0.001	-0.008	-0.006	-0.001
0.021	0.036	-0.027	0.000	-0.003	-0.006	0.001	-0.012	0.024	0.004	-0.016	-0.005	-0.023
-0.035	0.036	-0.057	-0.007	0.001	-0.120	0.012	0.008	0.029	-0.003	0.003	0.003	0.023
-0.031	0.035	-0.048	-0.004	0.002	-0.019	0.059	0.023	0.053	-0.062	-0.079	-0.015	-0.013
0.011	0.034	-0.021	-0.002	-0.004	-0.001	0.001	-0.008	0.016	0.001	-0.011	0.004	0.006
0.011	0.034	-0.022	-0.002	-0.003	-0.004	0.002	-0.002	-0.002	0.001	-0.004	0.006	0.006
-0.035	0.033	-0.007	-0.009	0.000	0.007	-0.022	-0.158	0.011	0.042	-0.033	-0.086	0.056
0.020	0.033	-0.030	-0.001	-0.004	-0.007	-0.003	-0.001	0.016	-0.001	-0.019	-0.018	0.027
0.017	0.033	-0.004	0.003	-0.002	-0.023	0.001	-0.002	-0.005	0.003	-0.002	0.002	-0.020
0.019	0.033	-0.028	0.001	-0.002	-0.006	-0.001	0.009	-0.010	-0.001	-0.007	-0.012	0.006
-0.047	0.032	-0.016	-0.009	-0.240	0.035	0.012	0.012	-0.017	-0.004	0.011	-0.070	0.030
0.010	0.032	-0.008	-0.001	-0.002	-0.002	0.002	0.000	-0.003	0.001	-0.003	-0.006	0.001
-0.028	0.032	-0.012	-0.006	0.005	0.003	0.012	0.000	-0.019	0.002	0.012	-0.060	-0.004
-0.040	0.031	0.007	0.005	0.014	-0.045	0.008	-0.029	-0.123	0.005	0.075	-0.007	0.034
0.020	0.031	-0.026	0.000	0.000	0.002	0.000	0.044	0.001	0.000	-0.016	-0.001	-0.025

Tabela 6.7 DBeta

Curva de Roc

Apresenta-se na tabela 6.8 a área da curva de ROC e a estatística de Kolmogorov-smirnov para os diferentes modelos estimados, subconjunto de treino e subconjunto de validação.

Modelo Logit	KS		ROC		Rácio de classificação global
	Treino	Teste	Treino	Teste	
60%-20%-20%	0,353	0,345	0,761	0,760	68,32%
70%- 15%-15%	0,349	0,360	0,764	0,764	69,04%
80%- 10%-10%	0,333	0,428	0,752	0,787	72,03%

Tabela 6.8 – avaliação da qualidade do modelo

O modelo que melhor generaliza os dados e, portanto, melhor responde ao objectivo do *credit scoring* é a representação de 80%-10%-10%.

Graficamente, as curvas de ROC apresentam o seguinte formato:

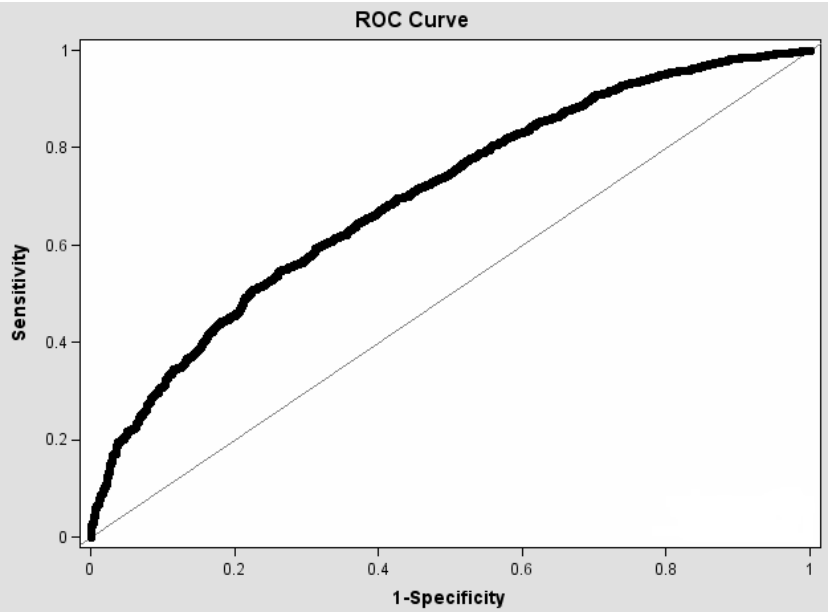


Figura 6.1 – Curva de ROC conjunto 80%-10%-10%.

6.2 Redes Neuronais

Iniciou-se a o processo de selecção da arquitetura da rede neuronal com as mesmas partições utilizadas na estimação do modelo logit 70%-15%-15% (conjunto de treino, validação e teste respectivamente) apurando para o efeito a estatística de RMSE. A rede com 11 neurónios foi a que apresentou um valor mais baixo. (apêndice B).

Seleccionamos a rede com 11 neurónios à luz daquela partição, recorremos ao software SAS (versão 9.1.3) módulo enterprise Miner para estimar os modelos, utilizando as mesmas amostras aleatórias utilizadas no modelo logit, sempre com a observância de proporções idênticas dos dois grupos considerados, regulares e *defaults*, a fim de melhor poder ajuizar da bondade das redes neuronais em classificar correctamente os indivíduos.

Foi considerado como função de activação a regressão logística e o método de estimação do gradiente com η=0.1 e μ=0.4. Subjugou-se a geração inicial dos pesos sinápticos à distribuição uniforme. As redes foram estimadas, tendo como função minimizadora a taxa de classificação errada.

À semelhança do que foi considerado para o modelo logit, também com as redes foi apurada qual a que apresentava melhor desempenho à curva ROC. A tabela 6.9 sumaria a comparação entre os modelos de redes neuronais consideradas utilizando o conjunto de validação na avaliação do modelo identificado.

Redes Neuronais	KS		ROC	
	Treino	Validação	Treino	Validação
60%-20%-20%	0,373	0,365	0,764	0,763
70%- 15%-15%	0,369	0,380	0,767	0,767
80%- 10%-10%	0,353	0,448	0,755	0,790

Tabela 6.9 –Curva ROC redes neuronais

É o conjunto 80%-10%-10% que apresenta um melhor desempenho na generalização do conjunto de dados.

Apresenta-se em apêndice B figura tradutora do gráfico de aprendizagem que se obtiveram na rede com 11 neurónios, para a melhor partição 80%-10%-10%, para o conjunto de treino.

Analisando o tradicional *cutoff* 50% associado à pontuação da probabilidade de incumprimento, pode se verificar igualmente que a rede neuronal apresenta um reduzido risco de crédito (erro tipo I).

Observados	Previsão	
	Regular	Default
Regular	76,06	21,64
Default	3,69	5,95

Tabela 6.10 Matrix de confusão

7 Conclusão:

No advento da entrada em vigor do acordo de Basileia II, as instituições financeiras munen-se de sofisticadas técnicas de análise de risco a fim de melhor optimizarem a sua carteira e, concomitantemente, a observância dos valores mínimos de capital requeridos para afectarem às diferentes carteiras de crédito.

O *credit scoring* aparece como o sistema usado pelas instituições financeiras para determinar a oportunidade de concessão de crédito a um solicitante. Levou-se a cabo assim, a tarefa de modelar uma base de dados real associada á carteira de crédito ao consumo (segemto Lar) de uma instituição Cabo-Verdiana, tendo sido competitvamente analisadas duas ferramentas utilizadas em *credit scoring*: O modelo logit (regressão logistica) e as Redes Neuronais.

As dezasete variáveis iniciais utilizadas para estudar o problema são as constantes na ficha de solicitação de crédito às quais se juntaram duas outras: o rácio de endividamento e a variável *target*. Relativamente à primeira, esta é defenida como o rácio entre o valor da prestação mensal e o rendimento do titular, representa uma importante variável de negócio que traduz a capacidade de um indivíduo fazer face ao serviço de divida que pretende contrair; já quanto à segunda, houve que recorrer à informação disponível nos planos de pagamentos a fim de possibilitar a sua classificação em regular ou em *default*, isto é, situação em que o cliente apresenta um atraso de 90 dias relativamente ao vencimento da prestação, numa base mensal. Assim todos os modelos foram elaborados, tendo como objectivo classificar operações de crédito de acordo com a sua probabilidade de apresentar atrasos no pagamento das prestações, tendo sido desprezados critérios de lucratividade que, conforme discutido constituiria uma abordagem alternativa a que propomos desenvolver.

Os indivíduos classificados como indeterminados foram expurgados da modelação de forma a garantir uma maior discriminação dos grupos, assim como os clientes inactivos. Optou-se por não excluir o conjunto dos indivíduos rejeitados a fim de possibilitar a avaliação do desempenho do modelo utilizando a técnica da inferência dos rejeitados como referido na página 41.

Debelada a classificação, seleccionou-se a janela de amostragem, período sobre o qual repousa a estimação dos modelos. Este período foi seleccionado após se ter verificado que o tempo médio de exposição dos contratos (tempo que media o momento do inicio do contrato e a data de obervação) se aproximava da maturidade da população (momento a partir da qual a taxa de incumprimento da carteira de crédito não evolui mais. Assim poderemos definir o cliente regular do *default* com maior segurança zelando pela qualidade da definição da variável *target.*

Para aferir a capacidade preditiva dos modelos estimados, subdividiu-se o conjunto amostral em três subconjuntos: subconjunto de treino, validação e teste. É sobre este penúltimo que se retiram as conclusões quanto à operacionalidade dos modelos de *credit scoring.*

A fim de possibilitar o melhor desenho do perfil dos clientes a amostra deverá encontrar-se balanceada. Dado o maior número de indivíduos em situação regular que em *default,* ponderaram-se os indivíduos regulares na percentagem que equilibrasse os dois grupos, emprestando desta feita uma maior riqueza na extracção dos ponderadores e, portanto, na melhor identificação de clientes.

Todas as variáveis disponibilizadas foram individualmente submetidas ao processo de categorização, que detectou grupos (categorias) de resposta homogênea em relação à varável *target*. Cabe referir que foram utilizados as mesmas variáveis para propósitos de comparação nas duas técnicas. Para ambas as ferramentes utilizou-se o *weight of evidence* (WoE) resultante do processo de categorização anteriomente referido como variáveis de *input.*

Para escolher a melhor arquitectura multicamada para as redes neuronais, nomeadamente a determinação do número de neurónios que deverão compor a camada intermédia, indiciado através do método de tentativa e erro, foi ensaiado a estatística de raíz quadrada do erro médio quadrático, tendo, esta apontado 11 neurónios como a rede que melhor discrimina as duas naturezas de indivíduos em estudo. As arquiteturas foram apuradas na exploração do conjunto 70%-15%-15%.

Geraram- se novos conjuntos aleatórios, 60%-20%-20% e 80%-10%-10% (identificando cada percentagem a dimensão do conjunto de treino, validação e teste respectivamente), a fim de poder conferir diferentes dimensões de parametrização às ferramentas envolvidas em comparação e garantir a capacidade de generalização que é fulcral nestes trabalhos.

O resultado do modelo logit e das redes neuronais foram comparados por recurso à curva de ROC.

Redes Neuronais	Logit	NN11
60%-20%-20%	0,760	0,763
70%- 15%-15%	0,764	0,767
80%- 10%-10%	0,787	0,790

Tabela 6.11- Comparação da área da curva de ROC

Quer no modelo logit, quer nos modelos de redes neuronais, é o conjunto de 80%-10%-10% aquele que sugere uma melhor generalização para os dados do subconjunto de validação.

A leitura das evidências numéricas associadas aos diferentes modelos ensaidos apontam ser o modelo baseado em redes neuronais como sendo o que melhor desempenho apresenta a prever o risco de crédito no mercado Cabo-veridano, quando comparado com o modelo logit. Contudo, para eleger estatisticamente o modelo que melhor se ajusta aos dados existentes em Cabo Verde, recorreu-se a estatístico U de Mann-Whitney proposto por (Delong E.R 1998) para comparar as áreas das diferentes curvas ROC associados aos diferentes modelos desenvolvidos, cujos resultados se apresentam na tabela 6.12.

Modelos	Chi-quadrado	d.f	Sig
Logit vs NN11	0,254	1	0,075

Tabela 6.12 - teste de DeLong e Clarke-Person (80%-10%-10%)

O ensaio do teste estatístico realizado permite concluir não haver evidência estatística a 95% de confiança para afirmar que as redes neuronais são preferíveis ao modelo logit (ou vice versa).

Apesar de todas as dificuldades técnicas e prácticas, dos modelos de *credit scoring*, esses modelos consistem em ferramentas bastante válidas para auxiliar o processo de análise de crédito, de uma forma obejctiva, racional e práctica, tendo em vista que o seu desempenho é sem dúvida superior aos métodos tradicionais (subjectivos) que ainda predominam em muitas instiutições em Cabo Verde.

Como análise global, considera-se que este processo de investigação, constituiu um valioso meio para discussão, e serviu para identificar, consolidar e sugerir linhas de investigação e abrir caminhos para o aprofundamento desta temática no seio dos investigadores Caboverdianos.

8 Limitações

A primeira dificuldade que surge em qualquer tarefa de modelação, mormente o *credit scoring*, diz respeito à elaboração de uma base de dados em condições apropriadas para o estudo. É preciso recolher e preparar um grande volume de dados, sendo necessário observar as condições de preechimento das variáveis e, caso necessário eliminar registos sobre os quais se desconfia da veracidade. A base de dados utilizada no presente estudo continha algumas variáveis com elevadas percentagens de *missing*, por outro lado não foi possível recolher muitas variáveis potencialmente discriminantes, (como por exemplo: Tipo de habitação, Antiguidade na habitação, antiguidade na profissão, número de dependentes, relação entre o primeiro e o segundo titular; informação do segundo titular... etc.). A ausência destas variáveis não prejudicou os modelos desenvolvidos, mas recomenda-se que sejam utilizados sempre que possível.

9 Bibliography

Amemiya, T. *Advanced Econometrics.* Oxford, 1985.

Ash, Dennis., e Steve Mester. *Best Practice in Reject Inferencing: Presentation at Credit Risk Modeling and Decisioning Conference.* Wharton FIC, University of Pensylvania, 2002.

Bação, Fernando Lucas. "Data Mining : Pós-Graduação em Estudos de Mercado e CRM." Lisboa, 2005.

Banasik, John, e Jonathan Crook. *Lean Models and Reject Inference.* Jounal of Operational Research Society, 2005.

Baptista, José Galvão. *O Custo de Intermediação Financeira em Cabo Verde -Factores Condicionantes* . Praia: Banco de Cabo Verde, 2006.

Barlett, P. *For valid generalization, the size of the weights is more important that the size of the network.* Advances in Neural Information Processing Systems, 9:134-140, 1997.

Beale, Jackson R.T. *Neuronal Computing: An introduction.* Adam Hilger Publishers. 1990.

Beck, N., G. King, e L. Zeng. *Improving Quantitative Studies of International Conflict:A Conjecture.* Vols. Vo1. 94, No. 1. American Political Science Review., 2000.

Blum, A. *Neural Networks in C++.* Vol. NY. Wiley, 1992.

Boger, Z., e H. Guterman. *Knowledge extraction from artificial neural network models.* Florida: IEEE Systems, Man, and Cybernetics Conference, 1997.

Boletim Económico. Praia: Banco de Cabo Verde, Fevereiro 2009.

Bose, N., e P. Liang. *Neural Network Fundamentals with Graphs, Algorithms and Applications.* USA: McGraw-Hill, 1996.

Braga, A. C. *Curvas ROC: Aspectos Funcionais e Aplicações :Tese de Doutoramento.* Braga: Universidade de Minho, 2000.

Braga, A. P., A. C. P. L. F. Carvalho, e T. B. Ludemir. *Redes Neurais Artificiais: Teoria e Aplicações.* Rio de Janeiro: LTC Livros Técnicos e Cientificos Editora S.A, 2000.

Burgo, Carlos. "Encontro de Governadores dos PALOP." *Encontro de Governadores dos PALOP.* Lisboa 19 e 20 de Setembro 2005: Banco de Cabo Verde, 2005.

Chorão, Luís António Ribeiro. *Logit vs Redes Neuronais Artificiais: Um exemplo aplicado a cartões de crédito.* Lisboa: Tese de Mestrado em Estatística e Gestão de Informação ISEGI-UNL, 2005.

Cloete, I. e J. M. Zurada. *Knowledje-based Neurocomputing.* Massachutsetts: Massachutsetts Institute of Technology, 2000.

Cortez, Paulo, e José Neves. *Redes Neuronais Artificiais.* Braga: Escola de Engenharia Universidade do Minho, 2000.

Crook, J. N., J. B. Banasik, e L. C. Thomas. *Sample Selection Bias in Credit Scoring Models.* Journal of the Operational Research Society,, 2003.

Crook, J., e J. Banasik. *Does Reject Inference Really Improve the Performance of Application Scoring Models?* Journal of Banking and Finance, 2004.

Damásio, A R. *O Erro de Decartes - Emoção, Razão e Cérebro Humano.* (D.Vicente e G.segurado,Tra 6ª ed): Publicação Europa -América, 1995.

Delong E.R, Delong D.M e D. Clarke-Pearson. "Comparing the Areas Under Correlated Receiver Operating Characteristics Curves: A Nonparametric Approach." *Biometrics (44), 837-845.*, 1998.

Dempster, P.A, N.M. Laird, e D.B. Rubin. *Maximum Likehood for imcomplete Data.* Journal of the Royal Statistics Society, 1977.

Efron, B., e R. Tibshirani. *An Introduction to the Bootstrap.* USA: Chapman & Hall, 1993.
Eisinga, R., P. Franses, e D. Dijk. *Timing of Vote Decision in First and Second Order Dutch Elections 1978-1995 Evidence from Artificial Neural Networks.* Oxford Journal, Political Analysis., 1997.

Feelders, A.J. *Credit Scoring and Reject Inference With Mixture Models.* Tilburg University, The Netherlands: Internation Journal of Intelegent System in Accounting, Finance and Management, 2000.

Freeman, James, e David M Skapura. *Neural Networks: Algorithms Applications and Programming Techniques.* Addison-Wesley Publishing, 1992.

Gallant, S. *Neural Network Learning and Expert Systems.* USA: MIT Press, Cambridge, 1993.

Gestel, Tony Van, e Bart Baesens. *Credit Risk Management: Basic concepts: Financial risk components, Rating analysis, models, economic and regulatory capital.* Oxford, 2009.

Gorni, A.A. *Redes Neuronais Artificiais - Uma abordagem revolucionária em inteligência artificial. Microsistemas,.* 1994.

Gurney, K. *An introduction to Neuronal Network.* London: UCL Press, 1997.

Hand, D.J., e W.E. Henley. *Can Reject Inference Ever Work?* IMA Journal of Mathematics Applied in Business and Industry, 1993.

Haykin, S. *Neuronal Networks - A Comprehensive Foundation.* New Jersey: Prentice Hall, 1999.

Henley, J. A, e B. J. McNeil. *The Meaning and Use of the Area Under the Receiver Operating Characteristics (ROC) Curve.* 1982.

Hosmer, David W, e Stanley Lemeshow. *Applied logistic regression.* Vol. Wiley series in probability and statistics. Texts and references section. New York: Wiley, 2000.

Hsai, D.C. *Credit Scoring and the Equal Credit Oportunity Act.* The Hasting Law Journal, 1978.

Joanes, D.N. *Reject Inference Applied to Logistic Regression for Credit Scoring.* IMA Journal of Mathematics Applied in Business and Industry, 1993.

Johnson, R.A., e D. W. Wichern. *Multivariate Statistics Analysis.* Vol. 5ª edição. New York: Printice Hall, 2002.

Kernsley, D., e T. Martinez. *A Survey Of Neural Network Research And Fielded Applications.* Vols. 2:123-133, 1992. International Journal of Neural Networks: Research and Applications.

Kohonen, T. *Self-Organizing Maps.* New York: Information Sciences, 2001.

Kosko, B. *Bidirectional Associative Memories.* Vols. SMC-18:49-60. IEEE Transactions on Systems, Man and Cybernetics, , 1988.

Kovacs, K. L. *Redes Neurais Artificiais - Fundamentos e Aplicações.* São Paulo: Editora Académica, 1996.

Kröse, B., e P. Smagt. *An Introduction to Neural Networks.* Vol. 8ª Edição. The University of Amesterdam, 1996.

Law, R., e R. Pine. *Tourism Demand Forecasting for the Tourism Industry:A Neural Network Network Approach.* In G. Peter Zang, Neural Networks in Businesses Forecasting. Chapter VI. IRM Press., 2004.

Levine, Ross. "Financial Development and growth: Schumpeter might be right." (Quarterly Journal of Economics) Vol. 108, no. 688-726. (1997).

Lewis, Edward M. *An Introduction to Credit Scoring.* Vol. Seconde Edition. San Rafael, California: Fair, Isaac and Co.,Inc., 1992.

Massoumi, E., A. Khotanzad, e A. Abay. "Artificial Neural Networks for Some Macroeconomic Series." *Econometric Reviews, 13(1).*, 1994.

Mateus, Abel. *Análise da eficiência e rentabilidade do sector bancário.* Praia Cabo Verde, 2000.

Mays, E. *Credit Scoring for risk managers: The Handbook for lenders.* Mason,OH, 2004.

Mays, Elizabeth. *Handbook of Credit Scoring.* Chicago: The Glenlake Publishing Company. Ltd, 2001.

McNelis, P. D. *Neural Networks in Finance:Gaining Predictive Edge in the Market.* Elsevier Academic Press., 2005.

Meneses, Maria Alexandrina da Silva. *As Redes Neuronais na Análise de Tráfico com o GPS:Dissertação de Mestrado em Posicionamento e Navegação por Satélite.* Faculdade de Ciências da Universidade do Porto. 2003.

Mester, Lorreta J. *What's the Point of Credit Scoring.* 1997.

Montrichard, Derek. *Reject Inference Methodologies in Credit Risk Modeling.* Toronto, Canada: Canadian Imperial Bank of Commerce, 2007.

Nargundkar, S., e J. Priestley. *Assessment of Evaluation Methods for Prediction and Classifications of Consumer Risk in the Credit Industry.* In G. Peter Zang,Neural Networks in Businesses Forecasting. Chapter XIV. IRM Press., 2004.

Neto, L.B. *Sistema híbrido de apoio à decisão para detecção e diagnóstico de falhas em redes elétricas. Dissertação de Mestrado em Engenharia Elétrica,.* Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 1997.

Neves, J. C., e A. Vieira. *Estimating Banruptcy Using Neural Networks Trained with Hidden Layer Learning Vector Quantization.* Lisboa: Working Paper,Departamento de Gestão, ISEG, UTL., 2004, Departamento de Gestão, ISEG, UTL.

Niu, Jack. "Managing Risks in Consumer Credit Industry." Beijing: Policy Conference on Chinese Consumer Credit, 2004.

Patterson, D. *Artificial Neural Networks - Theory and Applications.* Singapore: Prentice Hall, 1996.

Raymond, Anderson. *The Credit Scoring Tookit Theory and Practice for Retail Credit Risk Management and Decision Automation.* New York: OXFORD University Press Inc., New York, 2007.

Reed, R.D., e MarsII. *Neuronal Smithing:Supervised Learning in feedward Artificial Neuronal Network.* Cambridge, MIT, 1999.

Reichert, A.k., C.C Cho, e G. M. Wagner. *An Examination of the Conceptual Issues Involved in Developing Credit Scoring Models.* Journal of Business and Economic Statistics, 1983.

Riedmiller, M., e H. Braun. *A Direct Adaptative Method for Faster Backpropagation Learning: The RPROP Algorithm.* San Francisco, CA, USA: In Proceedings of the IEEE International Conference on Neural Networks, 1993.

Roisenberg, Mauro. *Emergência da Inteligência em Agentes Autónomos através de Modelos Inspirados na Natureza.* Florianópolis: Tese de Doutoramento em Engenharia Elétrica: Universidade Federal de Santa Catarina, 1998.

Roisenberg, Mauro, e Renato Corrêa Vieira. "Redes Neuronais Artificiais: Um Breve Tutorial."

Russel, S., e P. Norvig. *Artificial Intelligence - A Modern Approach.* New Jersey, USA: Prentice-Hall, 1995.

Sarle, W. *Neural network.* 1999.

Stopped Training and Other Remedies for Overfitting. In Proceedings of the 27th Symposium on the Interface of Computer Science and Statistics, pages 352-360,, 1995.

Sarmento, António. *Experimentação e avaliação de modelos para um problema de atribuição de Crédito:Tese de mestrado em análise de dados e sistemas de apoio à decisão.* Porto: Universidade do Porto Faculdade de Economia, 2005.

Schumpeter, Joseph. *The theory of Economic Development; traduzido por Redvers Opie, Cambridge,.* Harvard University Press , 1911.

Shachmurove, Y. *Applying Artificial Neural Networks to Business, Economics and Finance.* CARESS Working Papers: UCLA Department of Economics., 2002.

Shin, H.W., e So Young Sohn. *Reject inference in credit operations based on survival analysis.* Seoul, South Korea: Department of Computer Science and Industrial Systems Engineering, 2006.

Siddiqi, Naeem. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* New Jersey: Jonh Wiley & Sons, Inc, 2006.

Stanton, T.H. *Credit and Loan Scoring: Tools for Improved Management of Federal Credit Programs.* Baltimore MD: Center for the Study of American Government, John Hopkins University, 1999.

Swingler, K. *Applying Neural Networks: A Practical Guide.* London: Academic Press, 1996.

Tabachnick, B G, e L S Fidell. *Using Multivariate Statiscts.* Vol. 4ª edição. 2001.

Thawornwong, S., e D Enke. *Forecasting Stock Returns with Artificial Neural Networks.*

G. Peter Zang, Neural Networks in Businesses Forecasting,Chapter III, IRM Press., 2004.

Thomas, Lyn C. *Consumer Credit Models: Pricing, Profit and Portfolios.* New York: Oxford University Press Inc., 2009.

Thomas, Lyn C.,Edelman, David B., e Jonathan N. Crook. *Credit Scoring and Its Applications.* 2002.

Turner, Robin Varghese e Michael. "The Benefits of Wider Participation in Full-File Credit Reporting in Latin America and the Costs of the Status Quo." (Information Policy Institute) Março 2006: 2.

Wasserman, P. D. *Neural Computing: Theory and Practice.* New York., 1989.

Wynn, Helen McNab & Anthea. *Principles and Practice of Consumer Credit Risk Management.* Vol. 2nd edition. Institute of financial services, 2003.

Yeung, T. Kwork e D. *Constructive algorithms for structure learning in feedforward neural networks for regression problems::A survey. IEEE Transactions on Neural Networks.* Vols. 8(3):630-645. 1999.

Zhang, Y., Akkaladevi,S., Vachtsevanos,G., e T. Lin. *Granular neural web agents for stock prediction.* Soft Computing 6 (2002) 406 – 41. Springer-Verlag., 2002.

Apendices

Apendice A – Modelo logit com conjunto de treino de 80%

Tabela de estimativas dos parâmetros

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.99	0.03	803	<.0001
X1	1	0.96	0.10	98	<.0001
X2	1	0.85	0.07	139	<.0001
X3	1	0.67	0.22	28	<0.002
X4	1	0.62	0.12	27	<0.002
X5	1	-0.58	0.38	58	<0.002
X6	1	0.56	0.59	46	<0.002
X7	1	0.52	0.30	89	<0.002
X8	1	0.51	0.68	65	<0.002
X9	1	0.47	0.14	67	<0.002
X10	1	0.43	0.26	59	<0.002
X11	1	0.48	0.07	89	<0.002
X12	1	0.80	0.22	88	<0.002
X13	1	0.89	0.03	48	<0.002

Testing Global Null Hypothesis: BETA=0

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	525	13	<.0001
Score	491	13	<.0001
Wald	431	13	<.0001

Tabela de contingencia do Hosmer- Lemeshow

Partition for the Hosmer and Lemeshow Test					
		TARGET= 0		TARGET = 1	
Group	Total	Observed	Expected	Observed	Expected
1	483	197	201	286	282
2	483	295	285	188	198
3	483	327	314	156	169
4	483	323	334	160	149
5	483	347	351	136	132
6	483	365	369	118	114
7	483	392	386	91	97
8	483	396	403	87	80
9	483	423	422	60	61
10	479	447	447	32	32

Teste de Hosmer Lemeshow

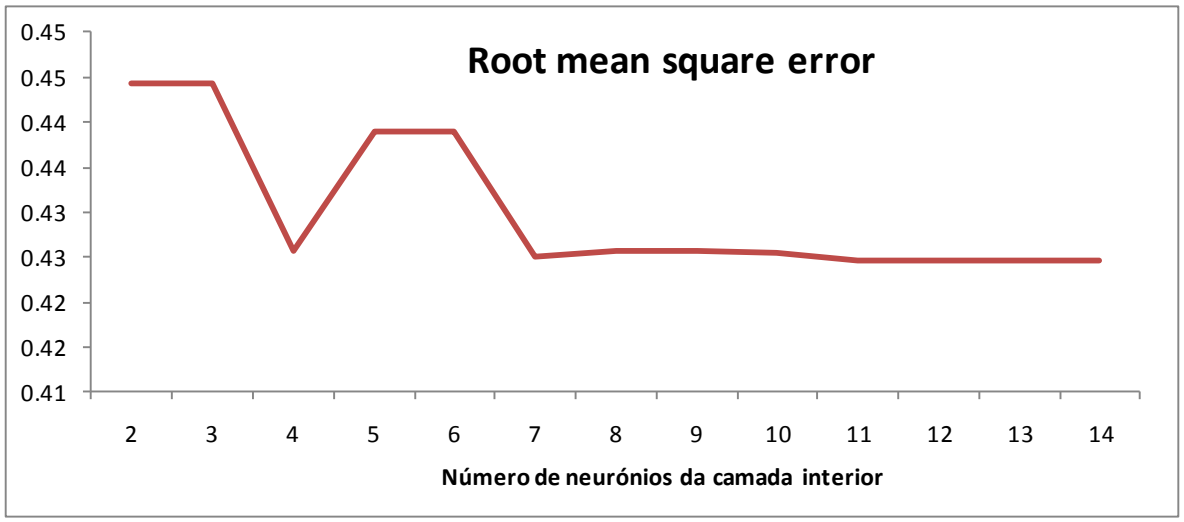
Test Hosmer and Lemeshow Goodness-of-Fit	
Chi-Square	Pr > ChiSq
5.2799	0.7273

Coeficiente de determinação – Pseudo R2

R-Square	0.17	Max-rescaled R-Square	0.19

Apendice B – Fit statistics RMSE

Selecção da melhor arquitetura RMSE



A rede com 11 neurónios é a que aprersenta menor erro.

Processo de Treino da rede Neuronal – conjunto 70%-15%-15%.